METRON

RIVISTA INTERNAZIONALE DI STATISTICA — REVUE INTERNATIONALE DE STATISTIQUE
INTERNATIONAL REVIEW OF STATISTICS — INTERNATIONALE STATISTISCHE ZEITSCHRIFT
REVISTA INTERNACIONAL DE ESTADISTICA

DIRETTORE PROPRIETARIO — DIRECTEUR ET PROPRIÉTAIRE
EDITOR AND PROPRIETOR — HERAUSGEBER UND EIGENTHÜMER
DIRECTOR Y PROPIEDARIO

Prof. Dott. Corrado Gini, *Direttore dell'Istituto di Statistica della Università di Roma*

AMMINISTRATORE — ADMINISTRATEUR — MANAGER — VERWALTER — ADMINISTRADOR

Dott. Franco Mariani, *Istituto di Statistica della Università di Roma*

SEGRETARI DI REDAZIONE — SECRÉTAIRES DE RÉDACTION
EDITORIAL SECRETARIES — REDACTIONSSECRETAERE — SECRETARIOS DE LA REDACCIÓN:

Proff. V. Castellano, N. Federici, G. Pompilj

Vol. XVI-N. 1-2 **15.VII-1951**

SOMMARIO-SOMMAIRE-CONTENTS-INHALT-SUMARIO

Corrado Gini. *Asimmetria e anormalità delle distribuzioni statistiche* . . . . pag. 3
Giuseppe Pompilj. *Sul problema inverso della ricerca statistica* . . . . . » 77
Fred C. Leone. *Some extensions of Bayes' theorem.* . . . . . » 107
Corrado Gini. *Della relazione tra differenza media con ripetizione e scostamento semplice medio dalla mediana* . . . . . » 123
Corrado Gini. *Del passaggio dall'indice di variabilità di un campione all'indice di variabilità della massa.* . . . . . » 127
G. Dahlberg. *Eine Methode zur Messung von Rassenunterschieden.* . . . . . » 133
Luigi L. Cavalli. *Un metodo rapido di calcolo della correlazione media fra più caratteri* . . . . . » 151
V. Castellano. *Sui quozienti di eliminazione.* . . . . . » 169
Tommaso Salvemini. *Su la determinazione dei quozienti di eliminazione* . . » 231
Karl Freudenberg. *Die Grenzen fur die Anwendbarkeit des Gesetzes der kleinen Zahlen* . . . . . » 285

ROMA

AMMINISTRAZIONE DEL METRON

UNIVERSITA' — ISTITUTO DI STATISTICA

Articoli giunti alla Rivista che verranno pubblicati nei prossimi numeri (*secondo l'ordine di arrivo*).
Articles received by the Review which will be published in future issues (*according to date of receipt*).
Articulos llegados a la Revista que se publicarán en los proximos numeros (*según el orden de su llegada*).
Articles reçus par la Revue et à paraître prochainement (*d'après la date de reception*).
Artikel die an die Zeitschrift angelangt sind und welche in den folgenden Nummern erscheinen werden (*nach der Reihenfolge des Eingangs*).

C. Gini. — *Probabilità che x termini di una serie erratica siano tutti decrescenti o tutti crescenti.*

C. Gini. — *Ulteriori considerazioni sull'asimmetria delle distribuzioni statistiche.*

C. Gini. — *L'evoluzione del concetto di media.*

E. Pizzetti. — *Dalle proporzioni continue alle progressioni.*

T. Salvemini. — *L'indice di dissomiglianza fra distribuzioni continue.*

Gli Autori degli articoli inviati per la pubblicazione nella Rivista, rinunciano in favore della medesima alla proprietà letteraria degli articoli stessi, qualora vengano pubblicati.
Les Auteurs des articles envoyés à la Revue, pour y être publiés, renoncent, en faveur de celle-ci, à la proprieté litteraire de leurs articles, s'ils sont acceptés.
The Authors of papers sent for publication in the Review are supposed to give up their copyright in favour of the Review if the papers are published.
Die Verfasser der zur Veröffentlichung in der Zeitschrift zugesandten Aufsätze, werden, falls selbige veröffentlicht werden, auf ihre Verfasserrechte zu Gusten der Zeitschrift verzichten mussen.
Los Autores de los artículos enviados para su publicación en « Metron » renuncian su propriedad a favor de la Revista cuando los artículos sean publicados.

CORRADO GINI

Asimmetria e anormalità delle distribuzioni statistiche[1]

Capitolo I.

Asimmetria e dissimetria delle distribuzioni statistiche.

1. È stato osservato che, mentre numerosi e raffinati indici sono stati proposti per misurare la variabilità, la concentrazione, ecc., dei fenomeni collettivi, non si posseggono ancora,

(1) Il presente articolo era già stato composto alla fine del 1940 e sarebbe dovuto essere pubblicato nel n. 2 del Vol. XIV del « Metron » destinato a comparire al principio del 1941. In una rassegna analitica, pubblicata in aprile 1941 (cfr. E. Pizzetti, *Recenti contributi alla metodologia statistica*, « Vita economica italiana », I trimestre 1941), il dott. Pizzetti, segretario di redazione del « Metron », riassumeva i principali risultati di questo articolo, che vennero così fin da allora resi noti al pubblico. (Cfr. nota a pag. 15). Se non che i risultati delle laboriose applicazioni dei metodi esposti, che erano state affidate allo stesso dott. Pizzetti, furono da questi in gran parte smarriti, e si dovettero rifare, mentre la sopravvenuta riduzione del formato della rivista impedì di includere l'articolo nel fascicolo del « Metron » a cui era destinato e che comprese i nn. 2-4 del Vol. XIV. Le complicazioni nazionali e internazionali che si sono susseguite hanno costretto a procrastinare la pubblicazione dei numeri successivi, mentre, dato il copioso materiale arrivato nel frattempo alla Rivista, parve opportuno, nella compilazione del vol. XV, di dare la precedenza ad altri articoli, così da assortire convenientemente argomenti, lingue ed autori. Queste circostanze spiegano il ritardo con cui l'articolo compare, in confronto all'anzidetta rassegna analitica. Del ritardo ho d'altronde profittato per estendere le applicazioni dei metodi esposti, le quali formeranno oggetto di un articolo successivo.

Alcuni risultati, con le formule relative e qualche esempio, furono frattanto esposti nel *Corso di Statistica* professato all'Università di Roma e sono riprodotti nella edizione del corso per l'anno accademico 1946-47 curata dai Proff. Vittorio Castellano, Nora Federici e Tommaso Salvemini ed edita, in italiano, dalla Casa Castellani - Roma 1947, ed in ispagnolo dalla Editorial Labor di Barcellona, 1951, e nella successiva, in corso di pubblicazione, per l'anno accademico 1950-51, curata dai proff. Giuseppe Pompilj e Nora Federici, ed edita dalla casa Veschi, Roma, 1951. Un riassunto dei

per questi, indici di asimmetria razionali (1). L'osservazione è fondata ed il fatto non resta senza spiegazione. Gli è che la misura della variabilità è necessaria per la soluzione di problemi biologici della massima importanza scientifica e pratica e la misura della concentrazione ricorre in problemi economici e finanziari vitali — ed altrettanto può dirsi per la misura della transvariazione, della connessione, della concordanza — mentre non vi sono questioni fondamentali che abbisognino, per la loro soluzione, di una misura dell'asimmetria delle distribuzioni statistiche. Aggiungasi che la misura dell'asimmetria di una distribuzione si riconduce immediatamente alla misura della dissomiglianza tra due distribuzioni e quest'ultimo problema è stato da tempo risoluto mediante gli indici di dissomiglianza (2), cosicchè chi degli indici di dissomiglianza sia al corrente può credere di disporre anche di una misura soddisfacente dell'asimmetria.

In realtà, però, la misura dell'asimmetria, anche se non è necessaria per risolvere problemi fondamentali, è utile, come si vedrà dalle applicazioni, per la soluzione di taluni problemi non privi d'importanza. Essa presenta, in ogni modo, un interesse per se stessa nel quadro della tecnica statistica, in quanto viene a completare il sistema di misure dei vari aspetti dei fenomeni collettivi, che costituiscono lo scopo primo dell'elaborazione statistica; nè è questo un'interesse puramente teorico, poichè solo eccezionalmente si danno curve che possono essere riguardate come perfettamente simmetriche, mentre non di rado l'asimmetria è pronunciatissima. D'altra parte, l'indice di dissomiglianza fornisce bensì la misura di un aspetto dell'asimmetria delle distribuzioni statistiche, ma questo aspetto non è il solo. Vi è infatti, come vedremo, un altro aspetto di tali distribuzioni che generalmente non viene distinto dal primo e che è invece opportuno distinguere anche con denominazioni diverse.

---

principali risultati trovasi pure esposto nella memoria *Metodologia statistica. La misura dei fenomeni collettivi* in « Enciclopedia delle Matematiche elementari », vol. III, Parte II, Milano, 1948, pp. 303-305.

(1) S. VIANELLI, *L'asimmetria dei fenomeni collettivi nella dinamica economica*, in G. PRONI - S. VIANELLI, *L'asimmetria statistica nella dinamica economica di alcune aziende agrarie*. « Annali dell'Osservatorio di Economia Corporativa per l'Emilia » Vol. VI, 1939, pag. 9.

(2) *Di una misura della dissomiglianza tra due gruppi di quantità e delle sue applicazioni allo studio delle relazioni statistiche*, « Atti del R. Istituto Veneto di Scienze, Lettere ed Arti », 1914.

Scopo di questo primo capitolo è di precisare le analogie e le differenze fra tali due aspetti, di indicare di ciascuno indici appropriati e di mettere in evidenza le relazioni di questi con altri indici già usati o proposti.

* * *

2. Cominciamo col dare la definizione precisa di *simmetria di una distribuzione.*

Il concetto di simmetria delle distribuzioni deriva evidentemente dalle loro rappresentazioni grafiche. Si suole dire che *una distribuzione è simmetrica quando, rappresentata graficamente, dà luogo ad una figura simmetrica.* Tale definizione è però insufficiente e, finchè non venga integrata, può risultare erronea; del che ci si persuade facilmente.

Anzitutto è da osservare che la stessa distribuzione può dar luogo ad una figura simmetrica od asimmetrica a seconda della categoria di curve con cui viene rappresentata [1]. Ad esempio, la distribuzione

0 1 2 3 4 5 6 7 8 9

dà luogo ad una curva di frequenza simmetrica rispetto alla linea *a b* (figura 1), ma ad una curva di concentrazione asimme-

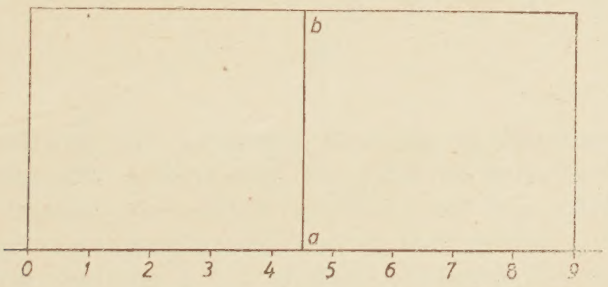


fig. 1

trica rispetto alla linea *c b* (figura 2). Si può dire, in generale, che tutte le distribuzioni, che corrispondono a curve di frequenza

---

[1] Noi distinguiamo tre categorie di curve di distribuzione: curve di graduazione, curve di frequenza, curve di concentrazione. Per le loro reciproche relazioni, cfr. *Intorno alle curve di concentrazione*, « Metron », Vol. IV, N. 3-4, 1932, pagg. 4-9.

simmetriche, corrispondono a curve di concentrazione asimmetriche [1], e viceversa che tutte le distribuzioni, le quali corrispondono a curve di concentrazione simmetriche, corrispondono a curve di frequenza asimmetriche.

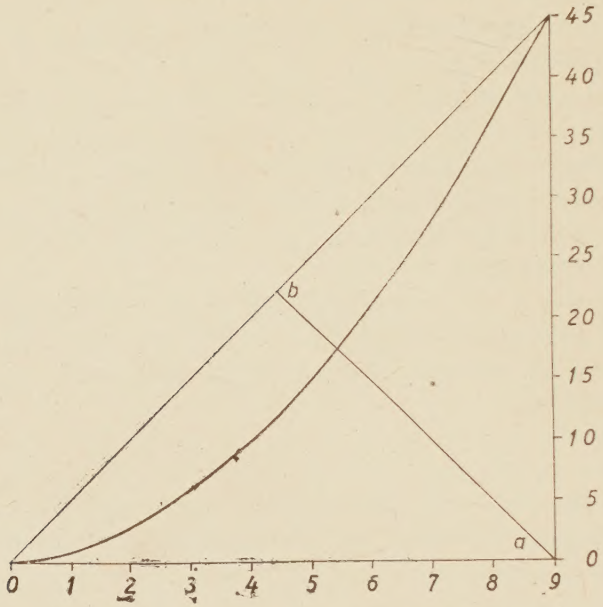


fig. 2

Conviene dunque precisare, anzitutto, che *una distribuzione simmetrica dà luogo ad una curva simmetrica di frequenza.*

Senonchè, non tutte le curve di frequenza simmetriche corrispondono a distribuzioni che noi riguardiamo come simmetriche. Le due curve di frequenza delle fig. 3 e 4 sono entrambe simmetriche, la prima rispetto alla linea *a b*, la seconda rispetto al punto *b*.

---

[1] Per curve simmetriche o asimmetriche, s'intendono curve che godono o non godono di *simmetria ortogonale*. Le curve di frequenza che godono di simmetria ortogonale danno luogo a curve di concentrazione che godono di *simmetria obliqua*. Cfr. V. CASTELLANO, *Sulle relazioni tra curve di frequenza e curve di concentrazione e sui rapporti di concentrazione corrispondenti a determinate distribuzioni*, « Metron », Vol. X, N. 4, 1933, pag. 44.

Esse corrispondono alle due seriazioni α e β, di cui in realtà noi, secondo il concetto corrente, diremmo simmetrica solo la α.

INTENSITÀ	NUMERO DEI CASI	
	Seriazione α	Seriazione β
1	1	1
2	2	2
3	4	4
4	5	5
5	4	6
6	2	8
7	1	9

Ci si persuade facilmente che le curve di frequenza simmetriche sono di due tipi: *tipo A*) curve simmetriche rispetto all'or-

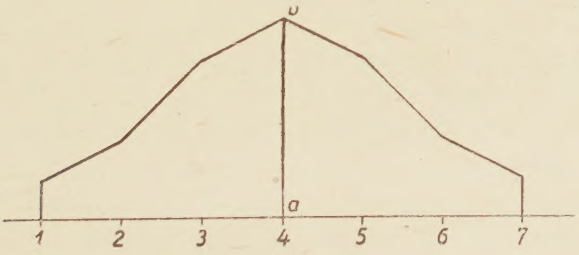


fig. 3

dinata mediana; *tipo B*) curve simmetriche rispetto all'apice di tale ordinata. Le curve di tipo B) si diranno *direttamente simmetriche*, in quanto i due rami della curva, a destra e a sinistra del punto *b*, possono farsi coincidere senza uscire dal piano, facendo rotare di 180° uno dei due rami con perno nel centro di simmetria *b*; le curve di tipo A) si diranno *inversamente simmetriche*, in quanto i due rami a destra e a sinistra dell'ordinata *a b* non possono farsi coincidere senza uscire dal piano; per ottenere la coincidenza, bisogna ripiegare il foglio, facendo asse della linea di simmetria *a b*.

È chiaro che noi riguardiamo come *simmetrica una distribuzione* solo *quando essa corrisponde* ad una curva di frequenza di tipo A), ossia ad una curva di frequenza simmetrica rispetto all'ordinata corrispondente alla semisomma degli estremi o, in altre parole, *ad una curva di frequenza inversamente simmetrica.*

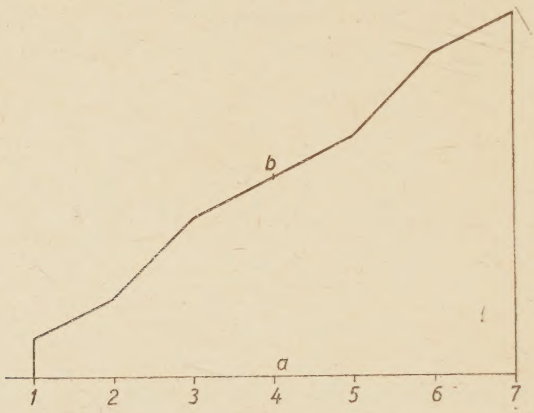


fig. 4

Si osservi che le curve di frequenza simmetriche di tipo A) corrispondono sempre a curve di graduazione simmetriche e, viceversa, le curve di graduazione simmetriche corrispondono sempre a curve di frequenza simmetriche di tipo A) (1), mentre le curve di frequenza simmetriche di tipo B) corrispondono a curve

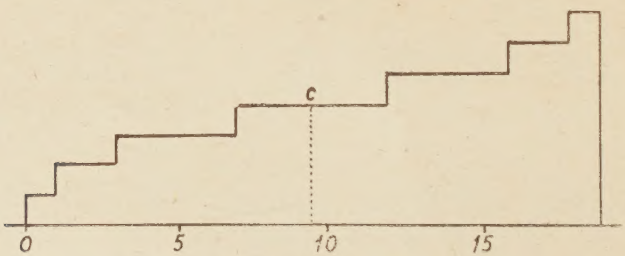


fig. 5

di graduazione asimmetriche. Così, la distribuzione α dà luogo alla curva di graduazione rappresentata nella figura 5, che è direttamente simmetrica rispetto al punto *c*, mentre la distribuzione

(1) Cfr. *Intorno alle curve di concentrazione*, op. cit., pagg. 7-8.

β dà luogo alla curva di graduazione nettamente asimmetrica rappresentata dalla figura 6.

Si noti pure che le curve di graduazione simmetriche godono sempre della simmetria diretta e sono tutte simmetriche rispetto al punto mediano.

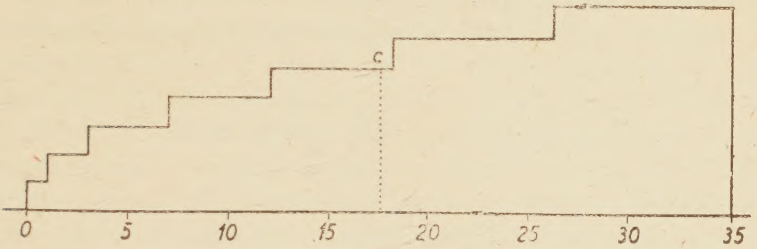


fig. 6

Potremo dunque definire come *simmetrica una distribuzione che dà luogo ad una curva di graduazione simmetrica.*

In luogo di tale definizione geometrica, potremo dare, della distribuzione simmetrica, una definizione analitica, dicendo che è *simmetrica una distribuzione quando risulta nullo l'indice di dissomiglianza tra scostamenti positivi e scostamenti negativi* (presi in valore assoluto) *dalla mediana.* L'indice di dissomiglianza tra due gruppi di quantità è dato, com'è noto, dalla media delle differenze assolute tra le quantità cograduate [1].

Se, avendo due gruppi di n quantità, si indica con a_i $(i = 1, 2, \ldots, n)$ la quantità generica i^{ma} di un gruppo e con b_i la quantità cograduata dell'altro gruppo, l'indice di dissomiglianza sarà

$$^1D = \frac{1}{n} \sum_{i=1}^{n} | a_i - b_i | .$$

Se il primo gruppo di quantità è composto dagli scostamenti positivi e il secondo dagli scostamenti negativi (presi in valore assoluto) dalla mediana di una distribuzione di m termini, sarà, per m pari, $n = \frac{m}{2}$, mentre, per m dispari, contando la

---

[1] Cfr. *Di una misura ecc.*, op. cit., pagg. 189-196.

mediana sia tra gli scostamenti positivi che tra i negativi, potremo porre $n = \frac{m+1}{2}$. Se poi con x_i $(i = 1, 2, \ldots, m)$ si indica il termine generico i^{esimo} della distribuzione e con M la mediana, e si suppone che sia gli m termini della distribuzione, sia gli n scostamenti positivi o negativi (presi questi ultimi in valore assoluto) siano ordinati in graduatoria decrescente si avrà $a_i = x_i - M$; $b_i = M - x_{m+1-i}$ e l'indice di dissomiglianza si potrà scrivere

$$^1I = \frac{1}{n} \sum_{i=1}^{n} | (x_i - M) - (M - x_{m+1-i}) | \qquad (1.1)$$

o anche

$$^1I = \frac{1}{n} \sum_{i=1}^{n} | x_i + x_{m+1-i} - 2M | \qquad (1.2)$$

Per che ^1I risulti nullo, è necessario e sufficiente che, ad ogni scostamento positivo dalla mediana della distribuzione, corrisponda uno scostamento negativo della stessa intensità; vale a dire che, nella curva di graduazione, ogni punto appartenente al ramo a destra del punto mediano sia simmetrico, rispetto a detto punto mediano, ad un punto appartenente al ramo a sinistra di detto punto mediano, condizione che rappresenta precisamente la condizione di simmetria delle curve di graduazione.

Le due definizioni, geometrica ed analitica, sopra date della simmetria di una distribuzione, risultano pertanto perfettamente equivalenti.

* * *

3. La definizione analitica della simmetria di una distribuzione ha il vantaggio di porgere immediato il passaggio alla misura dell'asimmetria.

Se, infatti, per che la distribuzione sia simmetrica, è necessario che sia $^1I = 0$, è evidente che, quanto più ^1I sarà elevato, tanto più si sarà lontani dalla condizione di simmetria e quindi tanto più pronunciata si dovrà dire l'asimmetria.

L'indice di dissomiglianza tra gli scostamenti positivi e gli scostamenti negativi (presi in valore assoluto) dalla mediana fornisce pertanto naturalmente e direttamente *la misura dell'asimmetria*

della distribuzione. Chiameremo pertanto ${}^1I$ *indice assoluto di asimmetria della distribuzione.*

L'asimmetria, così misurata, varia naturalmente secondo l'unità adottata per la misura del fenomeno, e la sua importanza è evidentemente maggiore o minore a seconda che la variabilità del fenomeno è più o meno forte.

Al fine di ottenere un indice di asimmetria della distribuzione che sia indipendente da tali circostanze, conviene, secondo il consueto procedimento adottato per gli indici statistici, di ragguagliare l'indice assoluto di dissomiglianza al suo valore massimo.

Il massimo dell'asimmetria non si può naturalmente determinare se non vincolato a certe condizioni, quale l'effettivo ammontare totale del carattere o la sua effettiva variabilità. Nel primo caso, si parla di *massimo assoluto* ; nel secondo caso, di *massimo relativo.*

Se non si pone altro vincolo se non quello che resti invariato l'ammontare totale del carattere, il massimo dell'asimmetria si avrà quando tutto l'ammontare del carattere, che indicheremo con T, sia concentrato in un solo termine. Si trova, in tale ipotesi,

$$Max\,({}^1I) = \frac{T}{n} = \frac{m}{n}\,A\,,$$

dove A indica la media aritmetica del carattere ed è $\frac{m}{n} = 2$ per m pari e $\frac{m+1}{n} = 2$ per m dispari. L'*indice relativo di asimmetria ragguagliato al massimo assoluto* risulta pertanto

$${}^1i' = \frac{1}{m\,A} \sum_{i=1}^{n} | x_i + x_{m+1-i} - 2\,M | \,. \qquad (2)$$

Se si pone, invece, il vincolo che resti invariata anche la variabilità del carattere, ci si potrà giovare, per stabilire il massimo di ${}^1I$, della considerazione che è

$$| (x_i - M) - (M - x_{m+1-i}) | \leqq (x_i - M) + (M - x_{m+1-i})\,,$$

(dove vale il segno $=$ quando uno dei due scostamenti cograduati è nullo) da cui

$$\sum_{i=1}^{n} |(x_i - M) - (M - x_{m+1-i})| \leqq m\,{}^1S_M\,,$$

dove ${}^1S_M$ indica lo scostamento semplice medio dalla mediana.

Nell'ultima relazione vale il segno = tutte le volte che almeno n delle m quantità sono uguali alla quantità massima o alla minima; esso vale in particolare nel caso di massima asimmetria, in cui tutto l'ammontare del carattere sia concentrato in un solo termine.

L'indice relativo di asimmetria ragguagliato al massimo relativo sarà pertanto.

$$ {}^1i'' = \frac{1}{m\,{}^1S_M} \sum_{i=1}^{n} | x_i + x_{m+1-i} - 2\,M | . \qquad (3.1) $$

Nel caso che tutto l'ammontare del carattere si trovi concentrato in un termine, risulta

$$ {}^1S_M = A $$

e quindi, come si richiede per la logica degli indici,

$$ {}^1i' = {}^1i'' . $$

In ogni altro caso è

$$ {}^1S_M < A $$

e quindi

$$ {}^1i' < {}^1i'' . $$

Oltre all'indice ${}^1D$ sopra considerato, detto *indice semplice di dissomiglianza*, è stato pure proposto ed usato l'*indice quadratico di dissomiglianza* (1), che risulta dalla media quadratica delle differenze tra le quantità cograduate ed è dato pertanto dalla espressione

$$ {}^2D = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (a_i - b_i)^2} . $$

---

(1) Cfr. *Nuovi contributi alla teoria delle relazioni statistiche*, « Atti del R. Istituto Veneto di Scienze, Lettere ed Arti », 1915, pagg. 1912-1914.

Se ne ricava, per *l'indice quadratico assoluto di asimmetria*, l'espressione

$$^2I = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i + x_{m+1-i} - 2\,M)^2} \qquad (4\,.\,1)$$

e, per *l'indice quadratico relativo di asimmetria*, secondo che si assuma per massimo il massimo assoluto o il relativo, le espressioni

$$^2i' = \frac{1}{^2A} \sqrt{\frac{1}{m} \sum_{i=1}^{n} (x_i + x_{m+1-i} - 2\,M)^2} \qquad (5)$$

$$^2i'' = \frac{1}{^2S_M} \sqrt{\frac{1}{m} \sum_{i=1}^{n} (x_i + x_{m+1-i} - 2\,M)^2} \qquad (6.1)$$

dove $^2A = \sqrt{\frac{T^2}{m}}$ indica la media quadratica della distribuzione nel caso in cui tutto l'ammontare del carattere sia concentrato in un solo termine, e ^2S_M indica lo scostamento quadratico medio della distribuzione dalla mediana.

* * *

4. Analoghi a quelli delle formule (1), (3), (4) e (6) sono gli indici proposti per misurare l'asimmetria da C. A. Bonferroni [1]. Il Bonferroni non mostrava di conoscere gli indici di dissomiglianza introdotti nella metodologia statistica una quindicina di anni prima e partiva da una definizione di asimmetria della distribuzione alquanto diversa, osservando che, nel caso di simmetria, le medie delle coppie di quantità cograduate (prima e ultima, seconda e penultima, terza e terzultima, ecc.) devono coincidere. Egli desumeva pertanto l'indice di asimmetria dalla variabilità di tali medie e perveniva alle formule

$$^1I = \sum_{i=1}^{m} | x_i + x_{m+1-i} - 2\,A | \qquad (7)$$

---

[1] *Elementi di statistica generale*, Anno acc. 1927-28, Torino, Gili, 1930, pag. 57 60; ristampa con aggiunte, Anno acc. 1932-33, Torino Gili, 1933, pagg. 60-64 e 27-28.

$$ {}_1i'' = \frac{\sum_{i=1}^{m} | x_i + x_{m+1-i} - 2A |}{2 \sum_{i=1}^{m} | x_i - A |} \qquad (8) $$

$$ {}_2I = \sum_{i=1}^{m} (x_i + x_{m+1-i} - 2A)^2 \qquad (9) $$

$$ {}_2i'' = \frac{\sum_{i=1}^{m} (x_i + x_{m+1-i} - 2A)^2}{2 \sum_{i=1}^{m} (x_i - A)^2} \qquad (10) $$

Si noti come il Bonferroni, da una parte, negli indici quadratici di asimmetria, consideri la somma dei quadrati degli scostamenti senza estrarne la radice quadrata — come si usa fare in Statistica nello stabilire gli indici quadratici (di variabilità, di connessione, di correlazione, ecc.) al fine di dare ai detti indici le stesse dimensioni dei rispettivi indici semplici — e come, d'altra parte, negli indici assoluti, consideri le somme degli scostamenti, anzichè le medie — come si usa fare in Statistica al fine di eliminare l'influenza del diverso numero di termini delle distribuzioni. Sono queste, però, imperfezioni facilmente rimediabili.

La relazione fra gli indici proposti dal Bonferroni e quelli a cui noi siamo pervenuti appare meglio mettendo le precedenti formule (1.2), (3), (4), (6) sotto la forma seguente:

$$ {}_1I = \frac{1}{2n} \sum_{i=1}^{m} | x_i + x_{m+1-i} - 2M | \qquad (1.3) $$

$$ {}_1i'' = \frac{\sum_{i=1}^{m} | x_i + x_{m+1-i} - 2M |}{2 \sum_{i=1}^{m} | x_i - M |} \qquad (3.2) $$

$$ {}_2I = \sqrt{\frac{1}{2n} \sum_{i=1}^{m} (x_i + x_{m+1-i} - 2M)^2} \qquad (4.2) $$

$$ {}_2i'' = \sqrt{\frac{\sum_{i=1}^{m} (x_i + x_{m+1-i} - 2M)^2}{2 \sum_{i=1}^{m} (x_i - M)^2}} \qquad (6.2) $$

che si desumono dalle rispettive (1.2), (3.1). (4.1), (6.1) tenendo presente che è

$$\sum_{i=1}^{m} | x_i + x_{m+1-i} - 2\,M |^k = 2 \sum_{i=1}^{n} | x_i + x_{m+1-i} - 2\,M |^k .$$

Appare allora che, a parte le imperfezioni sopra avvertite, gli indici proposti dal Bonferroni differiscono, da quelli a cui noi siamo pervenuti con le formule (1), (3), (4), (6), in quanto gli scostamenti, in luogo di essere misurati a partire dalla mediana, sono misurati a partire dalla media aritmetica [1].

Gli indici desunti dalle formule (8) e (10) hanno però il difetto di non raggiungere il valore 1 nel caso di massimo di asimmetria. In tal caso, che si verifica quando tutto l'ammontare del carattere sia concentrato in un solo termine, si ha infatti

$$Max\,(^1i'') = \frac{m-2}{m-1} = Max\,(^2i'') .$$

---

[1] Il Bonferroni, nelle edizioni del 1930 e 1933, avvertiva che l'indice semplice di asimmetria avrebbe potuto essere formato anche mediante gli scostamenti dalla mediana, nel quale caso — egli soggiungeva — l'indice vale 1 quando la mediana cade su uno dei due termini estremi, e mostrava poi come l'indice quadratico di asimmetria, formato mediante gli scostamenti dalla media, non potesse raggiungere il valore 1. Egli però non avvertiva il vantaggio che l'indice quadratico formato con gli scostamenti dalla mediana avrebbe presentato di fronte all'indice, da lui proposto, formato con gli scostamenti dalla media e neppure avvertiva che l'indice semplice, formato con gli scostamenti dalla media, non può raggiungere il valore 1, chè anzi dichiarava che esso è variabile da 0 a 1.

In una successiva edizione del suo corso (Torino, Gili) — che, a differenza delle precedenti, non porta data di pubblicazione, ma è riferita alle lezioni impartite all'Università commerciale Bocconi di Milano nello anno accademico 1940-41 — il Bonferroni ha modificato le anzidette formule, misurando gli scostamenti a partire dalla mediana (che dichiara dover essere centro di simmetria) e avvertendo che, quando essi sono misurati dalla media, nè l'indice semplice, nè l'indice quadratico raggiungono mai il valore 1 (cfr. pagg. 78-80). Non so se questa modificazione sia il frutto di una fortuita coincidenza o l'effetto della rassegna analitica del dott. Pizzetti, apparsa nella « Vita economica » del I trimestre 1941, in cui si esponeva il procedimento da me seguito per gli indici di asimmetria. Nell'edizione del 1933, in contrasto con la precedente del 1930, il Bonferroni parlava di « indici di dissimetria » anzichè di « indici di asimmetria », mentre nell'ultima ritorna alla terminologia di « indici di asimmetria », conforme anche in ciò alle proposte da me fatte in questo articolo (cfr. pagine 26-27) e ricordate nella anzidetta rassegna del Pizzetti.

Oltre a ciò, gli indici (1), (3), (4) e (6), a cui noi siamo pervenuti, hanno il vantaggio di ricollegarsi direttamente alla definizione di asimmetria adottata in geometria e all'indice di dissomiglianza, già in uso nella metodologia statistica e posto a base di altri indici statistici.

Definire come simmetrica una distribuzione che, rappresentata graficamente, dà luogo ad una curva di graduazione simmetrica, ci pare invero preferibile a definire come simmetrica (come fa il Bonferroni) una distribuzione che risulti da una serie di coppie di valori i cui centri di simmetria coincidono, e ciò non solo in quanto la prima definizione si ricollega direttamente alla definizione di simmetria adottata dai geometri, ma anche in quanto essa è più semplice della seconda. La prima definizione presuppone infatti un solo centro di simmetria, la mediana; la seconda presuppone $\frac{m}{2}$ o $\frac{m+1}{2}$ centri di simmetria.

Si aggiunga che, poichè, nella misura dell'asimmetria, si viene ad accoppiare ogni quantità con la sua cograduata, pare conveniente di assumere come centro di simmetria un valore, quale è il valore mediano, che sia interno a tutte le coppie, e da esso misurare gli scostamenti. Perciò, anche nel seguito di questa memoria, noi daremo sempre la preferenza agli indici basati sugli scostamenti dalla mediana, anzichè dalla media aritmetica o da altri valori medi.

* * *

5. Il concetto e la misura di asimmetria sopra esposti, direttamente derivati dal concetto di simmetria di una figura adottato in geometria, si basano sopra i valori assoluti delle differenze tra i termini cograduati o sopra i loro quadrati, a prescindere dal segno positivo o negativo delle dette differenze.

Si può, in tal caso, parlare di *asimmetria in senso assoluto* o *di asimmetria assoluta*.

In contrapposto a tale concetto, si parla, in Statistica, di *asimmetria positiva* o *negativa*, e gli indici adottati per misurarla possono naturalmente assumere valore positivo o negativo.

Tali indici sono sorti con particolare considerazione alle distribuzioni asimmetriche rappresentate dai poligoni di frequenza

binomiali o ipergeometrici ed alle curve che ne ha dedotto K. Pearson (1).

Si tratta di curve di frequenza asimmetriche unimodali che si possono considerare divise in due rami, uno a destra e uno a sinistra della moda, corrispondenti agli scostamenti positivi e rispettivamente negativi rispetto alla moda, dei quali rami uno ha una inclinazione meno forte e quindi si prolunga più dell'altro, dando luogo a scostamenti di maggiore intensità. In tal caso la media risulta superiore o inferiore alla moda a seconda che la curva di frequenza si prolunghi dalla parte degli scostamenti positivi o dei negativi e l'asimmetria viene rispettivamente detta positiva o negativa. La differenza

$$A - N, \tag{11}$$

dove N indica il valore della moda o *norma*, viene appunto assunta come *indice assoluto di asimmetria positivo o negativo* e il rapporto

$$\frac{A - N}{{}^2S_A}, \tag{12}$$

come *indice relativo di asimmetria positivo o negativo* (2). (S'intende come, in luogo dello scostamento quadratico medio dalla media aritmetica, si potrebbe assumere per denominatore un altro indice di variabilità). Siffatto indice di asimmetria risponde bene allo scopo per i tipi di curve di frequenza in considerazione delle quali è stato ideato, ma non per altri: manifestamente non è adatto per le curve di frequenza plurimodali.

D'altra parte, data la difficoltà pratica di determinare la moda per quelle distribuzioni che, per il numero limitato di casi

---

(1) Cfr. K. Pearson, *Contributions to the Mathematical Theory of Evolution*. II. *Skew Variation in Homogeneous Material*, « Phil. Trans. of the R. Soc. of London », Series A, Vol. 186, Part 1, 1895; X. *Supplement to a Memoir on Skew Variation*, Ibidem, Vol. 197, 1901; XIX. *Second Supplement to a Memoir on Skew Variation*, Ibidem, Vol. 216, 1916; « *Das Fehlergesetz und seine Verallgemeinerungen durch Fechner und Pearson* », *A Rejoinder*, « Biometrika », Vol. IV, Parts I and II, June 1905.

(2) Cfr. K. Pearson, *Skew Variation*, op. cit., pag. 370 e «*Das Fehlergesetz und seine Verallgemeinerungen durch Fechner und Pearson* », *A Rejoinder* art. cit. pagg. 173, 181.

a cui si riferiscono o per la presenza di forti cause perturbatrici, si presentano come affette da intense irregolarità, è parso opportuno di basare l'indice di asimmetria, anzichè sulla differenza tra media aritmetica e moda, sulla differenza tra media aritmetica e mediana, traendo partito dalla considerazione che, per le curve unimodali moderatamente asimmetriche (1), si può ritenere

$$A - N = 3\,(A - M).$$

Taluni autori, anzi, hanno adottato senz'altro, per indice assoluto di asimmetria, la differenza

$$A - M, \tag{13}$$

a cui fa riscontro l'indice relativo di asimmetria

$$\frac{A - M}{{}^2S_A}. \tag{14}$$

Ed è stato dimostrato (2) che è sempre

$$-1 < \frac{A - M}{{}^2S_A} < 1. \tag{14 \textit{bis}}$$

L'indice ricavato dalla (13) o dalla (14) si presenta quindi come un'approssimazione empirica di un indice valido solo per certi tipi di curve; ma in realtà, come vedremo in seguito (pag. 24), ad esso, con una leggera variante, si può attribuire un significato generale.

A seconda che la curva di frequenza unimodale si prolunga nel senso degli scostamenti positivi o dei negativi, la somma delle

(1) Cfr. K. Pearson, *Skew Variation*, op. cit. pag. 375 ; G. U. Yule, *Notes on the History of Pauperism in England and Wales, etc. Supplementary Note on the Determination of the Mode*, « Journ. Roy. Stat. Soc. », Vol. 59, 1896, pagg. 343 e segg. ; A. T. Doodson, *Relation of the Mode, Median and Mean in Frequency Curves*, « Biometrika », Vol. XI, part. IV, 1917, pagg. 425 e segg. ; G. U. Yule and M. G. Kendall, *An Introduction to the Theory of Statistics*, Eleventh edition, London, Griffin, 1937, pag. 125.

(2) H. Hotelling and L. M. Salomons, *The Limits of a Measure of Skewness*, « Annals of Mathematical Statistics », Vol. 3, N 2, May. 1932 ; R. Garver, *Concerning the limits of a Measure of Skewness*, Ibidem, Vol. III, N. 4, Nov. 1932. In realtà la dimostrazione di Hotelling e Salomons, come quelle di Garver, suppone m dispari, ma la disuguaglianza (14 bis) vale in ogni caso (cfr. nota a pag. 24).

potenze dispari degli scostamenti risulterà positiva o negativa. La media di tali potenze può perciò assumersi come indice assoluto di asimmetria positiva o negativa e il rapporto di tale media allo scostamento quadratico medio come indice relativo di asimmetria positiva o negativa (1).

Per che detto indice relativo sia un puro numero, indipendente, pertanto, dall'unità di misura delle quantità, è però necessario che il numeratore e il denominatore abbiano la stessa dimensione, ciò che si può ottenere ragguagliando la media potenziata (2) di ordine dispari degli scostamenti dalla mediana (${}^{2r+1}S_M$) o dalla media aritmetica (${}^{2r+1}S_A$) alla loro media potenziata di ordine pari (${}^{2s}S_M$ o rispettivamente ${}^{2s}S_A$).

Si ottengono così due categorie di indici relativi rappresentati dalle formule

$$\frac{{}^{2r+1}S_M}{{}^{2s}S_M} \qquad (15) \; ; \qquad \frac{{}^{2r+1}S_A}{{}^{2s}S_A} \qquad (16) \; .$$

Un caso particolare delle (16), per $s = r = 1$, è dato dalla formula

$$\beta = \frac{{}^{3}S_A}{{}^{2}S_A} \; . \qquad (16\ bis)$$

In particolare K. Pearson ha proposto, come indice assoluto di asimmetria il terzo momento della media aritmetica, ossia la media dei cubi degli scostamenti dalla media aritmetica

$$\mu_3 = {}^{3}S_A^3 \qquad (17)$$

e, come indice relativo, il rapporto

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{{}^{3}S_A^6}{{}^{2}S_A^3} = \beta \qquad (16\ ter)$$

---

(1) Cfr. le ripetute edizioni degli *Appunti di Statistica*, raccolti successivamente da vari studenti alle nostre lezioni, impartite dal 1914 in poi all'Università di Padova, prima, e, poi, a quella di Roma ; nonchè la traduzione in ispagnolo fattane dal prof. J. A. Vandellos sotto il titolo *Curso de Estadistica*, Editorial Labor, Barcellona, 1935, pagg. 209-210.

(2) Per il concetto di *media potenziata* e per altre categorie di medie, cfr. il nostro articolo *Di una formula comprensiva delle medie*, in «Metron», Vol. XIII, N. 2, febbraio 1938.

dove μ_2 è il secondo momento dalla media aritmetica, ossia la media dei quadrati degli scostamenti dalla media aritmetica (1).

Meglio però sarebbe assumere come indice relativo il rapporto

$$\frac{\mu_3}{\mu_2^{3/2}} = \beta^3$$

che risulta positivo o negativo a seconda che è positivo o negativo l'indice assoluto μ_3, mentre β_1 è sempre positivo e non ci può quindi dire se l'asimmetria è positiva o negativa, a meno che non si voglia convenire di dare a β_1 il segno positivo o negativo a seconda che è positivo o negativo μ_3. (2)

Il Pearson ha anche dimostrato che, chiamando μ_4 il quarto momento dalla media aritmetica, ossia la media aritmetica delle quarte potenze degli scostamenti dalla media aritmetica, e posto $\beta_2 = \frac{\mu_4}{\mu_2^2}$, si può ritenere, per una larga categoria di curve,

$$\frac{|A - N|}{{}^2S_A} = \frac{\sqrt{\beta_1}}{2} \cdot \frac{\beta_2 + 3}{5\,\beta_2 - 6\,\beta_1 - 9}. \qquad (18)$$

Il Pearson riteneva che si potesse trascurare il secondo fattore del secondo membro e quindi porre approssimativamente (3)

---

(1) Cfr., per l'indice assoluto μ_3, *Skew Variation*, op. cit. pag. 347 e «*Das Fehlergesetz und seine Verallgemeinerungen durch Fechner und Pearson*», *A Rejoindev*, art. cit., pag. 176. In quest'ultimo passo, è proposto l'indice relativo β_1.

(2) Anche G. M. KENDALL nel noto trattato *The Advanced Theory of Statistics* (vol. I, Second Edition, London, Griffin, 1945, p. 82), pubblicato, dopo che questo articolo era stato scritto (la prima edizione è del 1943) osserva che β^3 è per certi scopi, che d'altronde egli non specifica, più conveniente che β_1.

Deve tuttavia tenersi presente che con β^3 (come con β_1) si viene ad attribuire un'importanza esagerata agli scostamenti estremi, come è risultato dalle applicazioni eseguite alle distribuzioni delle stature (cfr., in proposito, la nostra comunicazione *La distribuzione delle stature è ipernormale*, presentata alla Società Italiana di Statistica nella seduta del 1° giugno 1949 (v. « Atti dell'VIII[a] riunione scientifica, 1-2 giugno 1949»). Un'edizione inglese, meno aggiornata, della comunicazione è stata pubblicata in «Acta Genetica et Statistica Medica», vol. I° fasc. 4, Uppsala, 1950. Per ciò che concerne l'accennato inconveniente dell'indice β^3, vedasi la nota (1) a pp. 40-41 dell'edizione italiana e a pp, 774-775 dell'edizione inglese.

(3) Cfr. G. U. YULE, *Notes on the History*, art. cit., pagg. 343-344, e K. PEARSON, «*Das Fehlergesetz und seine Verallgemeinerungen durch Fechner und Pearson*», *A Rejoinder*, art. cit., pagg. 177 e 181.

$$\frac{|A - N|}{{}^2S_A} = \frac{\sqrt{\beta_1}}{2} \qquad (18\ bis)$$

Nelle formule (18) e (18 bis) abbiamo scritto $|A-N|$, perchè, il valore di β_1 essendo sempre positivo, deve prendersi come positivo, salvo contrarie convenzioni, anche $\sqrt{\beta_1}$: effettivamente Yule e Kendall insegnano che la misura data dalla (18) deve sempre prendersi senza segno (1), ed altrettanto dovrebbe dirsi per le misure date dalla (18 bis). Se non che in realtà non vi è nessuna ragione di rinunciare a rappresentare anche il segno delle differenze $A-N$, in quanto, in realtà, la dimostrazione del Pearson porta alla relazione

$$\frac{A - N}{{}^2S_A} = \frac{\mu_3}{2\,\mu_2^{3/2}} \frac{\beta_2 + 3}{5\,\beta_2 - 6\beta_1 - 9} \qquad (19)$$

da cui, adottando l'approssimazione suggerita dallo stesso Pearson,

$$\frac{A - N}{{}^2S_A} = \frac{\mu_3}{2\,\mu_2^{3/2}} \qquad (19\ bis)$$

e anche.

$$A - N = \frac{1}{2} \frac{\mu_3}{\mu_2^3}\,. \qquad (19\ ter)$$

Le formule (19) e (19 *bis*) dovrebbero dunque servire in pratica per determinare l'indice relativo di asimmetria, positiva o negativa, rappresentato dalla (12) e la (19 *ter*) per determinare approssimativamente l'indice assoluto rappresentato dalla (11), evitando le difficoltà della determinazione diretta del valore modale; in realtà, però, esse non sono adoperate, ciò che, per la (19), dipende verosimilmente dalla sua complicazione e, per la (19 *bis*) e (19 *ter*) non so se si debba attribuire ad essere le formule poco conosciute, oppure dall'essersi riconosciuto che la loro approssimazione è lungi dall'essere soddisfacente (2).

Un altro gruppo di indici di asimmetria positiva o negativa si basa sulla differenza fra certi scostamenti contrograduati dalla mediana presi in valore assoluto, per esempio tra gli scostamenti corrispondenti al primo e al terzo quartile oppure al secondo e all'ottavo decile, e via dicendo.

---

(1) Cfr. YULE and KENDALL. *Op. cit.*, pag. 163.

(2) Effettivamente non trovo riportata la (19 *bis*) nei trattati di statistica, che ho consultato; neppure in quello dello Yule, che pure ne aveva fornito applicazioni (cfr. nota 3 a pagina precedente).

È particolarmente usato l'indice desunto dalla differenza tra gli scostamenti dalla mediana corrispondenti al primo quartile Q_1 e al terzo quartile Q_3. L'espressione

$$(Q_3 - M) - (M - Q_1) = Q_3 + Q_1 - 2M \tag{20}$$

dà in tal caso l'indice assoluto di asimmetria positiva o negativa, mentre, come indice relativo, è conveniente prendere l'espressione

$$\frac{(Q_3 - M) - (M - Q_1)}{(Q_3 - M) + (M - Q_1)} = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1} \tag{21}$$

che ha per limite ± 1 [1].

Presenta una certa analogia formale con quest'ultimo indice, ma ne è, in realtà — come vedremo in appresso — sostanzialmente diverso, un indice di asimmetria positiva o negativa proposto dal Vianelli e dato dall'espressione

$$\frac{D_d - D_s}{D_d + D_s} \tag{22}$$

dove D_d indica la differenza media tra gli scostamenti positivi e D_s la differenza media tra gli scostamenti negativi dalla mediana [2].

L'indice del Vianelli è basato, come l'autore mette debitamente in luce, su una convenzione, di cui avremo occasione di esaminare più innanzi il fondamento.

* * *

6. Poichè si ha simmetria, o invece asimmetria, in senso assoluto, quando gli scostamenti dalla mediana di un segno sono uguali, o rispettivamente diversi, dagli scostamenti cograduati di segno contrario, e poichè, in senso assoluto, l'asimmetria è tanto più elevata quanto più forti sono tali differenze tra scostamenti cograduati, viene naturale di dire che vi è asimmetria positiva, o viceversa negativa, a seconda che sieno superiori le differenze positive o viceversa le negative tra gli scostamenti cograduati, vale a dire a seconda che sia

$$\sum_{i=1}^{n} (x_i - M) \gtrless \sum_{i=1}^{n} | x_{m+1-i} - M |$$

---

[1] Cfr. YULE and KENDALL, op. cit., pag. 162.

[2] Op. cit., pag. 23.

e che l'asimmetria positiva o negativa è tanto più forte quanto più tale disuguaglianza è pronunciata.

Questo ovvio criterio porta ad adottare come indici semplici di asimmetria positiva o negativa le stesse formule già adottate per gli indici semplici di asimmetria assoluta, con la sola differenza che al numeratore, invece di considerare la somma delle differenze assolute tra scostamenti cograduati, si considera la somma algebrica delle loro differenze.

L'indice semplice assoluto di asimmetria positiva o negativa può pertanto scriversi

$$^1I_p = \frac{1}{n} \sum_{i=1}^{n} (x_i + x_{m+1-i} - 2\,M) =$$

$$= \frac{1}{n} \left(\sum_{i=1}^{n} (x_i + x_{m+1-i}) - 2\,n\,M' \right) \qquad (23.1)$$

e gli indici relativi corrispondenti

$$^1i_p' = \frac{1}{m\,A} \left(\sum_{i=1}^{n} (x_i + x_{m+1-i}) - 2\,n\,M' \right) \qquad (24.1)$$

$$^1i_p'' = \frac{1}{m\,{}^1S_M} \left(\sum_{i=1}^{n} (x_i + x_{m+1-i}) - 2\,n\,M' \right). \qquad (25.1)$$

Nel caso di m pari, è

$$m = 2\,n\,;\; 2\,n\,M = m\,M\,;\; \sum_{i=1}^{n} (x_i + x_{m+1-i}) = m\,A$$

e, nel caso di m dispari,

$$m = 2\,n - 1\,;\; 2\,n\,M = (m+1)\,M\,;\; \sum_{i=1}^{n} (x_i + x_{m+1-i}) = m\,A + M\,.$$

Per m pari o dispari, il numeratore degli indici si può quindi porre sotto la forma

$$m\,(A - M)$$

e quindi gli indici assumono le forme seguenti:

$$^1I_p = \frac{m}{n}\,(A - M) \qquad (23.2)$$

$$ {}^1i_p' = \frac{A - M}{A} \qquad (24.2) $$

$$ {}^1i_p'' = \frac{A - M}{{}^1S_M} , \qquad (25.2) $$

dei quali l'ultimo differisce dall'indice $\frac{A - M}{{}^2S_A}$ (formula (14)) solo per la sostituzione al denominatore dello scostamento semplice medio dalla mediana allo scostamento quadratico medio dalla media aritmetica (1).

Diremo dunque che vi è *asimmetria positiva o negativa a seconda che sia* $A \gtrless M$.

Quando è $A = M$ saremo tratti naturalmente a dire che vi è *asimmetria nulla della distribuzione*. Senonchè non è detto affatto che in tal caso vi sia *simmetria* della distribuzione. Nelle figure 7 e 8, ad esempio, la linea continua rappresenta una distribuzione asimmetrica in cui è $A = M$; si avrebbe invece una distribuzione simmetrica se il secondo ramo della curva di graduazione nella fig. 7 o il primo nella fig. 8 seguisse la linea tratteggiata.

Evidentemente è contraddittorio, dal punto di vista logico, ammettere, da una parte, che vi sia asimmetria nulla, cioè che non vi sia asimmetria, e, dall'altra, che non vi sia simmetria. Si è quindi condotti necessariamente alla conclusione che, nella misura dell'asimmetria assoluta e nella misura della cosidetta asimmetria positiva o negativa, si attribuiscono alla parola asimmetria significati diversi che in realtà sarebbe opportuno distinguere con termini diversi.

---

(1) Il valore di ${}^1S_M$ è sempre inferiore al valore di ${}^2S_A$, essendo sempre, per una nota proprietà della mediana, ${}^1S_M < {}^1S_A$ e, per una nota relazione tra media aritmetica e media quadratica, ${}^1S_A < {}^2S_A$. Se, come si è visto (pagg. 11-12) ${}^1S_M$ rappresenta il massimo di $A - M$, ${}^2S_A$ non lo può rappresentare Effettivamente, Hotelling e Salomons hanno provato che, per m dispari. il valore di $\frac{A - M}{{}^2S_A}$ è inferiore a + 1 e superiore a — 1 (cfr. pag. 18), ma tali limiti non possono essere mai raggiunti. Infatti, per che potessero essere raggiunti, tutti gli scostamenti dovrebbero essere uguali tra loro, nel qual caso sarebbe ${}^2S_A = {}^1S_A = {}^1S_M$; ma sarebbe allora $A - M = 0$ e quindi il valore dell'indice risulterebbe nullo,

* * *

7. Si verifica, nel campo della simmetria delle distribuzioni, quello che si è verificato nel campo delle relazioni statistiche.

Sotto la parola « correlazione » andavano un tempo confusi due concetti essenzialmente diversi : il concetto di connessione e quello di concordanza.

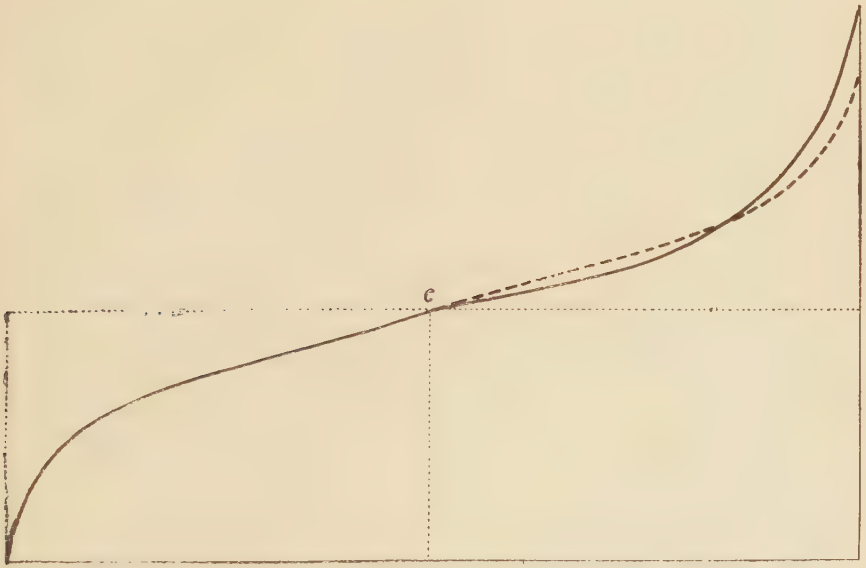


fig. 7

Il concetto di connessione implica la dipendenza della distribuzione delle modalità di una variabile dalle modalità di un'altra variabile senza tener conto del verso secondo cui la dipendenza si avvera ; introducendo (quando ciò sia possibile) anche la nozione del verso, si passa dal concetto di connessione a quello di concordanza. Quando non vi è connessione, non vi è neppure concordanza ; ma può non esservi concordanza pur essendovi connessione (1).

---

(1) Cfr. il nostro articolo *Sur le coefficient de corrélation comme caractéristique du dégré de la concordance* pubblicato nella « Revue de l'Inst. Int. de Stat. », 1936, 3ème livraison (pag. 500 e segg.) e le memorie, in esso citate, apparse negli « Atti dell'Istituto Veneto di Scienze, Lettere ed Arti » negli anni 1914-1917.

L'aver debitamente distinto i concetti di connessione e concordanza ha fatto compiere un notevole progresso alla teoria delle relazioni statistiche.

Analogamente, il concetto di asimmetria in senso assoluto implica la esistenza di differenze tra gli scostamenti contrograduati dalla mediana, indipendentemente dal verso secondo cui le differenze si verificano; introducendo (e ciò in questo caso è sempre possibile) la nozione di verso, si passa al cosidetto concetto di asimmetria positiva o negativa. Quando non vi è asimmetria in senso assoluto, non vi può essere evidentemente asimmetria in senso positivo o negativo, ma vi può non essere asimmetria in senso positivo o negativo, pur essendovi asimmetria in senso assoluto.

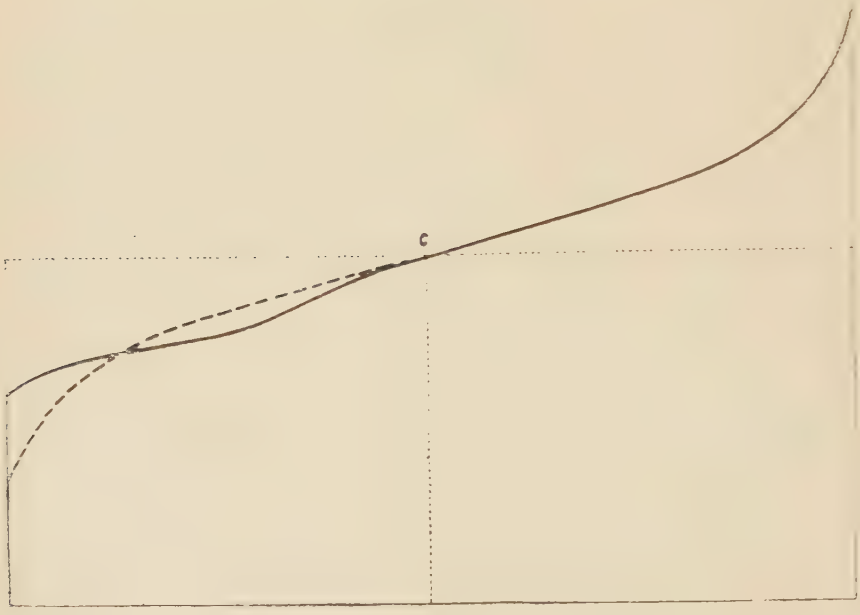


fig. 8

Come nel campo delle relazioni statistiche, così nel campo dell'asimmetria, sarà opportuno distinguere con parole diverse i due concetti. Io proporrei di serbare all'« asimmetria in senso assoluto » il vocabolo *asimmetria*, già usato in tal senso dai mate-

matici, e di parlare, invece che di « asimmetria positiva o negativa », di *dissimmetria* o di *scentramento*. La parola « dissimmetria », avrebbe il vantaggio di richiamare l'analogia di concetto con l'asimmetria, ma starebbe a indicare un'asimmetria che implica uno squilibrio ; la parola « scentramento » starebbe a indicare, non solo l'esistenza di tale squilibrio, ma anche la sua natura. La caratteristica della dissimmetria è, infatti, che la media aritmetica, che è il centro di gravità, non coincide con la mediana che è il centro di simmetria della curva di graduazione.

Si dirà che *una figura presenta dissimmetria positiva se il centro di gravità è spostato nel senso dei valori positivi rispetto al centro di simmetria* : si dirà, invece, che *una figura presenta dissimmetria negativa se il suo centro di gravità è spostato nel senso dei valori negativi*. Tali figure si possono pure dire *scentrate* ; e *scentramento*, rispettivamente *positivo* o *negativo*, può opportunamente chiamarsi il fenomeno che in esse si verifica.

Quando non vi è asimmetria, non vi è neppure dissimmetria, nè positiva nè negativa ; il centro di gravità della figura coincide col suo centro di simmetria ; può dirsi pertanto che la figura è *centrata*, ossia che vi è *scentramento nullo*.

* * *

8. Per che la mediana coincida con la media aritmetica è necessario e sufficiente che la media aritmetica degli scostamenti positivi dalla mediana sia uguale alla media aritmetica degli scostamenti negativi presi in valore assoluto. Indicando con S_p e S_n tali medie la (23.1)o (23.2) può infatti mettersi anche sotto la forma

$$ {}^1I_p = {}^1S_p - {}^1S_n \qquad (23.3) $$

e la (25,1)o (25.2) sotto la forma

$$ {}^1i_p'' = \frac{{}^1S_p - {}^1S_n}{{}^1S_p + {}^1S_n} \qquad (25.3) $$

la quale mostra chiaramente le relazioni dell'indice relativo di dissimmetria, a cui noi siamo logicamente pervenuti, con quello,

già in uso, basato sopra i quartili e con quello proposto recentemente dal Vianelli. Il nostro indice ragguaglia alla loro somma la differenza tra le medie aritmetiche di due distribuzioni di scostamenti, mentre l'indice basato sui quartili ragguaglia alla loro somma la differenza tra le loro rispettive mediane e quello proposto dal Vianelli ragguaglia alla loro somma la differenza tra i rispettivi indici di variabilità, desunti dalle differenze medie.

Poichè è ben noto che due distribuzioni possono avere la stessa variabilità e medie diverse e, viceversa, possono avere la stessa media aritmetica e variabilità differenti, è chiaro che l'indice proposto dal Vianelli non può fornire una misura corretta della dissimmetria o scentramento, nel senso sopra definito. Nel caso in cui due distribuzioni siano di tipo analogo, potrà risalirsi più o meno approssimativamente dalla differenza tra i due indici di variabilità alla differenza tra le due medie, ma non in casi diversi, Il Vianelli deduce precisamente il suo indice dall'ipotesi (che egli vorrebbe elevare a convenzione) che le somme delle quantità contrograduate della distribuzione

$$x_m + x_1 \,,\; x_{m-1} + x_2 \,, \ldots$$

tendano a crescere quando l'asimmetria è negativa e a diminuire quando è positiva. Ora, tale ipotesi corrisponde al vero quando le distribuzioni degli scostamenti positivi e dei negativi sieno dello stesso tipo, ma può non corrispondere in altri casi.

La media aritmetica non è che un caso particolare della media potenziata di grado r^{mo} per $r = 1$. L'indice fornito dalle (23.3) e (25.3) può pertanto considerarsi come un caso particolare di una classe di indici che si basano sulla differenza tra le due medie potenziate di grado r^{mo} — relative l'una agli scostamenti positivi e l'altra agli scostamenti negativi (presi in valore assoluto) — ragguagliata alla somma di dette medie. Se con ${}^rS_p$, ${}^rS_n$ si indicano le dette medie potenziate di grado r^{mo}, l'indice assoluto di dissimmetria sarà

$$ {}^rI_p = {}^rS_p - {}^rS_n \tag{26}$$

e l'indice relativo

$$ {}^ri_p'' = \frac{{}^rS_p - {}^rS_n}{{}^rS_p + {}^rS_n} \tag{27}$$

che, per $r = 1$, si riducono alla (23.3) e rispettivamente alla (25.3).

Si potrebbe pensare di dare invece, degli indici generalizzati, le formule seguenti

$$ {}^{r}I_p = \sqrt[r]{{}^{r}S_p{}^{r} - {}^{r}S_p{}^{r}} \qquad (28.1) $$

$$ {}^{r}i_p'' = \sqrt[r]{\frac{{}^{r}S_p{}^{r} - {}^{r}S_n{}^{r}}{{}^{r}S_p{}^{r} + {}^{r}S_n{}^{r}}} \qquad (29.1) $$

che possono scriversi anche

$$ {}^{r}I_p = \sqrt[r]{\frac{1}{n} \sum_{i=1}^{n} [(x_i - M)^r - (M - x_{m-1-i})^r]} \qquad (28.2) $$

$$ {}^{r}i_p'' = \sqrt[r]{\frac{\sum_{i=1}^{n} [(x_i - M)^r - (M - x_{m+1-i})^r]}{\sum_{i=1}^{n} [(x_i - M)^r + (M - x_{m+1-i})^r]}} \qquad (29.2) $$

e per $r = 1$ si riducono pure alle (23.3) e (25.3).

Per r dispari, la (28.2) diviene equivalente alla media potenziata di grado r^{mo} degli scostamenti e da tale equivalenza si intendono le analogie e le differenze che intercedono tra gli indici del tipo (29) e quelli del tipo (15). Le due categorie di indici, per r dispari, hanno uguale il numeratore e diverso il denominatore e gli indici del tipo (29) hanno il vantaggio di restare tra i limiti $+1$ e -1, ciò che non può dirsi degli indici del tipo (15).

* * *

9. L'asimmetria e la dissimmetria (o scentramento) sono dunque due aspetti diversi delle distribuzioni statistiche. Di essi ciascuno ha la sua importanza ed il suo campo di applicazione. Non si può quindi convenire nella proposta del Vianelli di abbandonare, ai fini dell'analisi statistica, il primo concetto, in uso tra i matematici, per attenersi al secondo concetto, che permette di distinguere il carattere positivo o negativo dell'asimmetria [1].

Se, conoscendo la forma di una certa distribuzione teorica, per esempio gaussiana, noi vogliamo giudicare di quanto una distribuzione effettiva se ne discosta per effetto del numero ristretto delle osservazioni o dell'intervento di fattori di perturbazione, è alla misura dell'asimmetria che dovremo ricorrere; se, invece, vogliamo giudicare se e di quanto, tra le cause pertur-

[1] Cfr. op. cit., pag. 25.

batrici sistematiche che sappiamo intervenire (per esempio, eliminazione dei valori più elevati o dei più bassi, quale si verifica per le stature nell'esame della idoneità dei coscritti a certe armi), prevalgono quelle che affettano gli scostamenti positivi o quelle che affettano i negativi, è la misura della dissimmetria che dobbiamo calcolare.

Vano e fallace sarebbe, d'altra parte, pretendere che uno stesso indice servisse a dare la misura di due aspetti diversi dei fenomeni collettivi quali l'asimmetria e la dissimmetria, e infondata appare pertanto la critica mossa dal Bonferroni ([1]), seguito dal Vinci ([7]) e dal Vianelli ([3]), ai consueti indici di dissimmetria usati in Statistica, in quanto questi possono risultare nulli quando non vi è asimmetria nulla. Tale critica è anche meno comprensibile da parte del Vianelli, in quanto egli distingue correttamente l'asimmetria dalla dissimmetria e propone egli stesso un nuovo indice di dissimmetria al quale potrebbero muoversi le stesse obbiezioni da lui mosse agli indici generalmente in uso. Siffatte critiche fanno riscontro alle critiche, ugualmente infondate, mosse al coefficiente di correlazione, in quanto non serve a misurare la connessione ([4]). Anche in questo, vi è stato parallelismo fra la teoria della connessione e della concordanza, da un lato, e la teoria dell'asimmetria e della dissimmetria, dall'altro.

([1]) Op. cit., pagg. 64-67.

([2]) *Manuale di Statistica*, Bologna, Zanichelli, vol. I, 1936, pagina 118-120.

([3]) Op. cit., pag. 13 e segg.

([4]) La questione fu anche portata davanti all'Istituto Internazionale di Statistica, che nominò un'apposita commissione per studiarla. Dopo una comunicazione e due rapporti di M. Fréchet (cfr. « Bull. dell'Ist. Int. de Statistique », Tome XXVIII, Session de Londres, 1934, 1ère livraison, pagg. 83-84, 2ème Livraison, pag. 25 e segg., 45 e segg.; e Tome XXIX, Session d'Athènes, 1936, 1ère livraison, pagg. 130-132 et 3ère livraison, pag. 318 e segg.) e vari articoli per parte del Fréchet stesso (Revue de l'Inst. Inter. de Stat. », 1935, 4ème livraison, pag. 365 e segg.), mia (Art. cit. *Ibidem*, 1936, 3ème livraison, pag. 355 e segg.) e del Pietra (*Ibidem*, 1936, 4ème livraison, pag. 500 e segg.), l'Istituto, accettando la mia tesi, approvò all'unanimità una mozione, in cui si constata che il coefficiente di correlazione non può misurare, salvo il caso di relazione lineare, il grado della *connessione* o dipendenza, ma può, indipendentemente dalla forma della relazione statistica, misurare il grado della *concordanza*, nella quale si considera, oltre che l'intensità, anche la direzione della dipendenza (cfr. Bull. cit., Tome XXIX, 1ère livraison, pagg. 130-131).

Capitolo II.

Anormalità, disnormalità e ipernormalità delle distribuzioni statistiche.

10. Un altro capitolo arretrato della metodologia statistica è quello che tratta della conformazione delle distribuzioni statistiche in confronto con la distribuzione gaussiana o *normale*. Qui non si è fatto alcun passo avanti, e neppure — che io sappia — alcun tentativo, dopo la trattazione che oltre mezzo secolo fa ne ha fatto K. Pearson (1).

K. Pearson ha chiamato *curtosi* tale aspetto della conformazione delle curve, distinguendo le curve in *platicurtiche* e *leptocurtiche*, a seconda che, a pari variabilità, si differenziano dalle curve normali per una forma più tozza o, viceversa, più snella, a seconda, cioè, che sono meno frequenti, in confronto alla curva normale, i valori centrali ed estremi o viceversa i valori intermedi. La fig. 9 dà, con la linea continua, una curva di frequenza platicurtica e la fig. 10 una curva di frequenza leptocurtica confrontate con una curva normale (rappresentata dalla linea tratteggiata).

Per avere la misura della curtosi, il Pearson partiva dalla quantità

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \qquad (30)$$

dove μ_2 e μ_4 indicano il secondo e rispettivamente il quarto momento intorno alla media aritmetica, ossia la media delle seconde o rispettivamente delle quarte potenze degli scostamenti dalla media aritmetica. Poichè, per la curva normale, è $\beta_2 = 3$ la quantità

$$\gamma_2 = \beta_2 - 3 \qquad (30\ bis)$$

(1) Cfr., per tutto questo, K. Pearson, *Contributions to the mathematical theory of evolution*, « Phil. Trans. of the R. Soc. of London », Series A, Vol. 185, 1894, particolarmente a pagg. 93-95 e « *Das Fehlergesetz*, ecc. », op. cit., 1905, particolarmente a pagg. 173-175 e 181.

veniva assunta dal Pearson come misura della curtosi. In base a questo criterio, una curva o la rispettiva distribuzione si dirà pertanto platicurtica o leptocurtica a seconda che sia

$$\beta_2 \lesseqgtr 3 .$$

Nel caso in cui sia $\beta_2 = 3$ la curva viene dal Pearson chiamata *mesocurtica*.

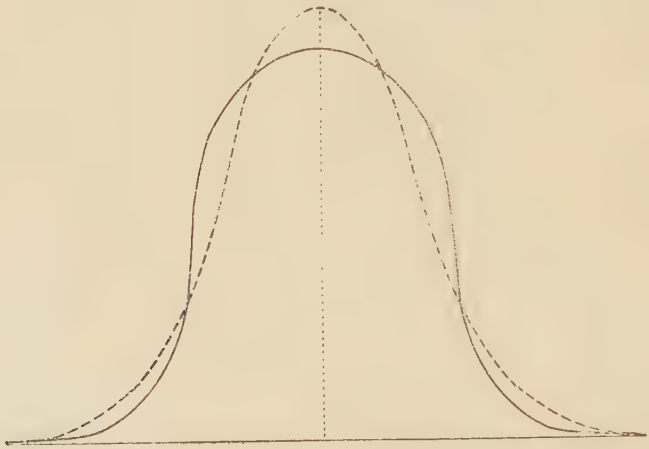


fig. 9

La misura della curtosi era stata dapprima ideata dal Pearson per misurare la divergenza dalla conformazione normale delle curve simmetriche[1] ; ma in seguito venne applicata indipendentemente dal carattere simmetrico o asimmetrico delle curve. L'asimmetria e la curtosi si possono effettivamente concepire come due aspetti indipendenti delle distribuzioni statistiche.

Il metodo di K. Pearson può venire generalizzato considerando come misura della curtosi la differenza tra il valore teorico ed il valore effettivo del rapporto tra due momenti di ordine diverso o tra due medie potenziate di diverso grado degli scostamenti. Nè è necessario che i momenti o le potenze siano di ordine pari quando si abbia l'avvertenza di prendere gli scostamenti negativi per il loro valore assoluto.

---

[1] Cfr. *Contributions* ecc., op. cit., pagg. 94-95.

Nel metodo così generalizzato, rientrano la differenza tra il rapporto dello scostamento quadratico medio allo scostamento semplice medio dalla media aritmetica ed il suo valore teorico in una distribuzione normale

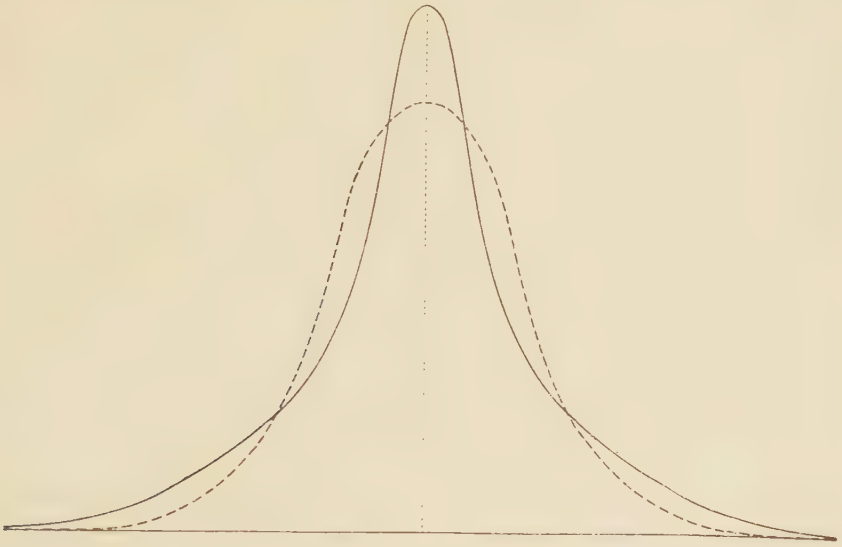


fig. 10

$$\frac{^2S_A}{^1S_A} - 1{,}2533 \qquad (31)$$

e la differenza tra il quadrato di tale rapporto e il suo valore teorico

$$\frac{^2S^2_A}{^1S^2_A} - 1{,}5708 \qquad (32.1)$$

o il doppio di tale differenza

$$2\,\frac{^2S^2_A}{^1S^2_A} - \pi\,. \qquad (32.2)$$

(1) Di tale indice vennero fatte ampie applicazioni nella già citata comunicazione *La distribuzione delle stature é ipernormale*. Esso presenta il vantaggio, non trascurabile nella pratica, che le operazioni, che il suo calcolo implica, si prestano facilmente al controllo. Può pertanto essere consigliabile adottarlo quando il calcolo ne è affidato a studenti o ad altro personale di cui è necessario controllare le operazioni.

Questi criteri furono anzi usati, molto prima che dal Pearson, da vari autori che si occuparono della teoria della dispersione, i quali, però, non sempre ne davano un'interpretazione corretta, ammettendo che la concordanza del valore effettivo col valore teorico stesse ad indicare una dispersione normale della distribuzione, mentre, in realtà, essa ne attesta solo la conformazione normale, che è cosa essenziamente diversa (1).

Gli indici suddetti sono dunque atti a dirci se una distribuzione è platicurtica, mesocurtica o leptocurtica ed anche se essa è più o meno platicurtica o leptocurtica, ma, non essendo ragguagliati ad un massimo, non si prestano a confrontare tra loro la curtosi di curve diverse, nè ci permettono di farci un'idea della portata da attribuire ad una certa differenza tra il valore teorico ed il valore effettivo del rapporto tra due momenti di ordine diverso o tra due loro funzioni. Supponiamo di aver trovato che il quadrato del rapporto dello scostamento quadratico medio allo scostamento semplice medio risulta per una distribuzione = 4, con un eccesso di 2,4292 in confronto al suo valore teorico = 1,5708. La differenza può essere giudicata forte; mentre, se usassimo del rapporto tra scostamento quadratico medio e scostamento semplice medio, otterremmo un valore = 2, la cui differenza di 0,7467, può essere, rispetto al valore teorico 1,2533, giudicata moderata. Effettivamente, la seconda differenza è solo del 60 % del suo valore teorico, mentre la prima sale al 160 %. Eppure si tratta della stessa distribuzione! D'altra parte, gli indici in parola si basano su criteri empirici e si limitano a dare la misura di un aspetto della divergenza tra le curve osservate e le curve normali.

Nelle pagine che seguono, ci proponiamo di riprendere la questione in forma più generale e di collegarne la soluzione alla misura dell'asimmetria e della dispersione.

* * *

11. Ritorniamo perciò agli indici di dissomiglianza.

La curtosi non è evidentemente che uno degli aspetti della dissomiglianza tra una curva e la curva normale.

(1) Cfr. *Il sesso dal punto di vista statistico*, op. cit., pagg. 80-131. La differenza tra *dispersione normale* e *conformazione normale* di una distribuzione è stata chiaramente posta in luce in tale volume, dove anzi, per evitare l'equivoco che era sorto per il doppio significato attribuito alla parola *normale*, proponevo di parlare di *dispersione regolare* quando la conformazione della curva era gaussiana (cfr. pag. 84).

Chiameremo tale dissomiglianza *anormalità*. Senonchè è necessario, a tal fine, definire la curva normale a cui la curva osservata si paragona. Per definire una curva normale è necessario conoscere un suo valore medio (e, in questo caso particolare, la mediana, la media aritmetica e la moda coincidono) ed un suo indice di variabilità (per esempio, lo scostamento quadratico medio o lo scostamento semplice medio).

Potremo dire pertanto che *una distribuzione* (o la rispettiva curva) *è anormale quando differisce dalla distribuzione normale*

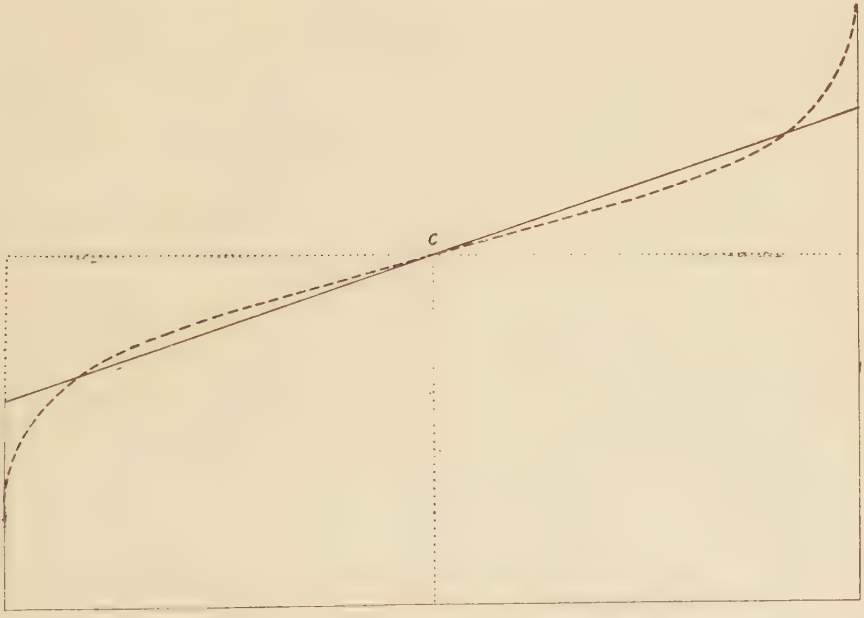


fig. 11

che ha la stessa mediana e la stessa variabilità rispetto alla mediana e che *essa è tanto più anormale quanto maggiore è la dissomiglianza tra essa e la curva normale.*

Nelle figure 11, 12, 13, 14, 15, 16 la linea continua dà la curva di graduazione corrispondente alla curva anormale, mentre la linea tratteggiata dà la rispettiva curva di graduazione normale od *ogiva*.

A misura della dissomiglianza, potremo assumere l'indice semplice o l'indice quadratico di dissomiglianza. Coerentemente, dovremo confrontare la distribuzione osservata con la distribu-

zione normale avente lo stesso scostamento semplice medio o rispettivamente lo stesso scostamento quadratico medio. Otterremo così due *indici assoluti di anormalità* ; l'*indice assoluto semplice di anormalità*, desunto dall'espressione

$$^1N = \frac{1}{m} \sum_{i=1}^{m} | x_i - y_i | \qquad (33)$$

e l'*indice assoluto quadratico di anormalità*, desunto dall'espressione

$$^2N = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (x_i - y_i)^2} \qquad (34)$$

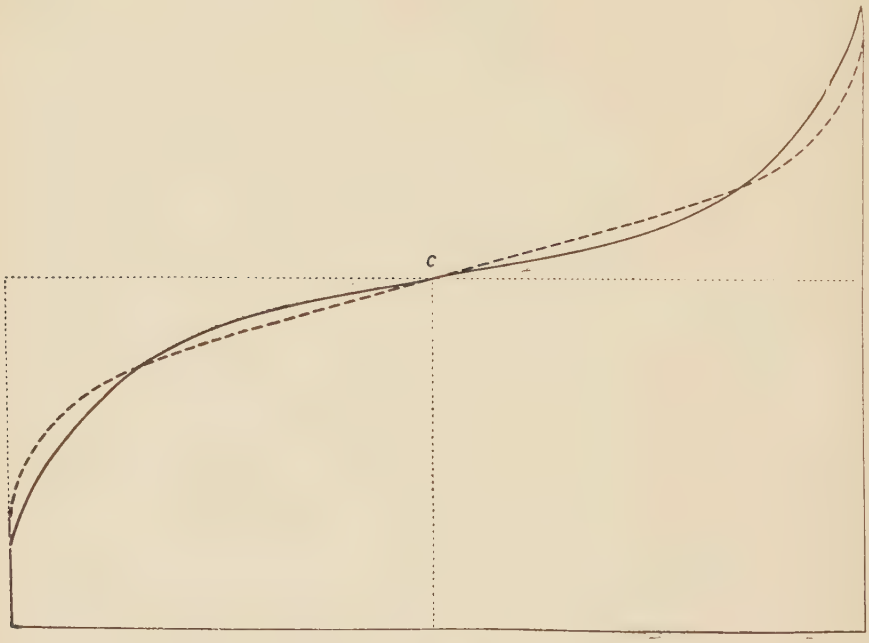


fig. 12

dove $i = 1, 2, \ldots\ldots, m$ indica uno degli m termini della distribuzione, x_i indica il valore dell'i^{mo} termine della distribuzione osservata ed y_i il valore cograduato della distribuzione normale.

Oltre agli indici di anormalità ^1N, ^2N, che sono indici assoluti, potremo costruire gli *indici di anormalità relativi* che si ottengono ragguagliando i precedenti ai loro massimi.

Ora il massimo di ^1N si può porre uguale al doppio dello scostamento semplice medio dalla mediana ed il massimo di ^2N si può

porre uguale allo scostamento quadratico medio dalla mediana moltiplicato per $\sqrt{2}$. (1)

L'*indice relativo semplice di anormalità* sarà pertanto fornito dall'espressione (2)

$$ {}^1n = \frac{{}^1N}{2\,{}^1S_M} = \frac{\sum_{i=1}^{m} | x_i - y_i |}{2 \sum_{i=1}^{m} | x_i - M |} \qquad (35) $$

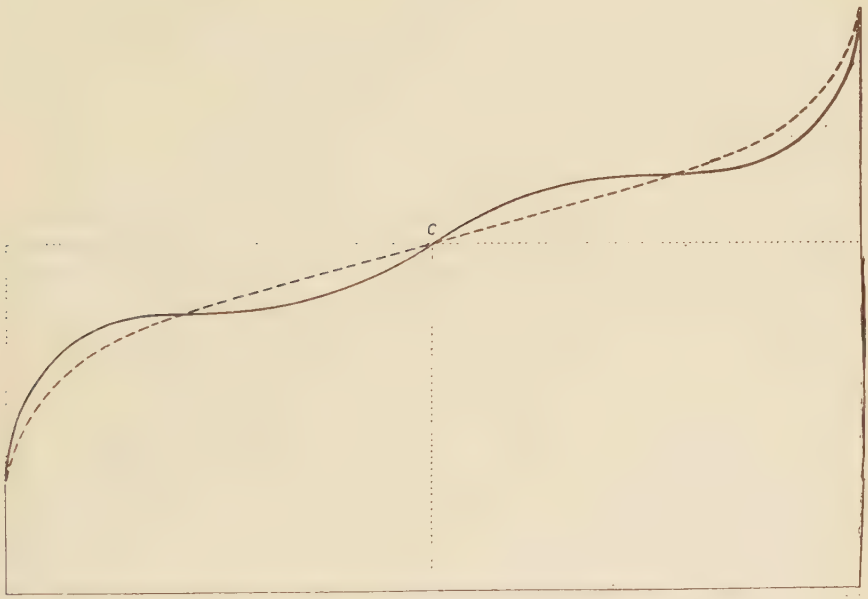


fig. 13

(1) Dalle disuguaglianze :

$$ | (x_i - M) - (y_i - M) | \leq | x_i - M | + | y_i - M | \text{ ossia} $$
$$ | x_i - y_i | \leq | x_i - M | + | y_i - M | $$
$$ [(x_i - M) - (y_i - M)]^2 \leq (x_i - M)^2 + (y_i - M)^2 \text{ ossia} $$
$$ (x_i - y_i)^2 \leq (x_i - M)^2 + (y_i - M)^2 $$

(di cui la seconda è giustificata dal fatto che i valori x_i e y_i sono cograduati e quindi i due scostamenti $x_i - M$ e $y_i - M$ hanno ugual segno) si deduce che il valore di ${}^1N$ non può mai superare il doppio dello scostamento semplice medio dell'indice, nè il valore di ${}^2N$ può mai superare il valore dello scostamento quadratico medio moltiplicato per $\sqrt{2}$. Ci si avvicina a tale massimo quando, degli m valori della distribuzione, $m-1$ sono uguali alla mediana e l'm^{mo} valore è uguale a m volte lo scostamento semplice medio o rispettivamente lo scostamento quadratico medio dalla mediana.

(2) Nell'edizione del 1947 del nostro *Corso di Statistica* e nella memoria inserita nell'« Enciclopedia delle Matematiche elementari » (cfr. nota 1 a pagina 3) venne per errore omesso, al denominatore, il fattore 2.

e *l'indice relativo quadratico di anormalità* dall'espressione (1)

$$^2n = \frac{^2N}{\sqrt{2}\,{}^2S_M} = \sqrt{\frac{\sum_{i=1}^{m}(x_i - y_i)^2}{2\sum_{i=1}^{m}(x_i - M)^2}}. \qquad (36)$$

È chiaro che gli indici semplici e quadratici, assoluti e relativi, di anormalità non sono vincolati a nessuna conformazione particolare della curva osservata, ma sono validi per curve unimodali o plurimodali, simmetriche o asimmetriche.

* * *

12. La trattazione precedente è perfettamente analoga a quella che ci ha condotto alla determinazione degli indici semplici e quadratici di asimmetria assoluta e relativa.

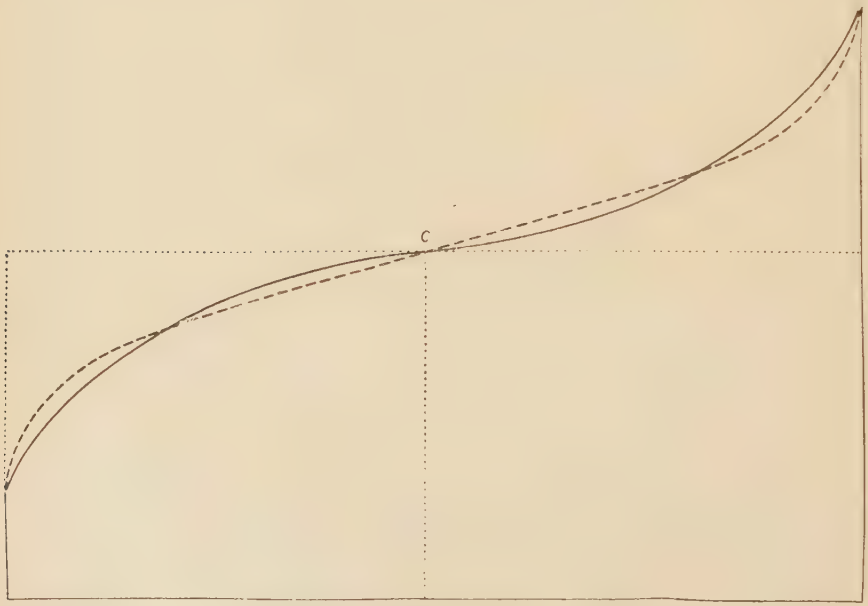


fig. 14

(1) Nell'edizione del 1947 del nostro *Corso di Statistica* venne per errore omesso, al denominatore, il fattore $\sqrt{2}$.

Analogamente a quanto si è fatto nel caso dell'asimmetria e della dissimmetria, noi potremo, pure in questo caso, prendere in considerazione, oltre che l'intensità, anche il verso delle differenze che intercedono tra le quantità cograduate della curva osservata e della curva normale ; ed anche qui sarà bene distinguere questo ulteriore aspetto delle distribuzioni da quello, sopra considerato, dell'anormalità. Al concetto di anormalità sopra esposto,

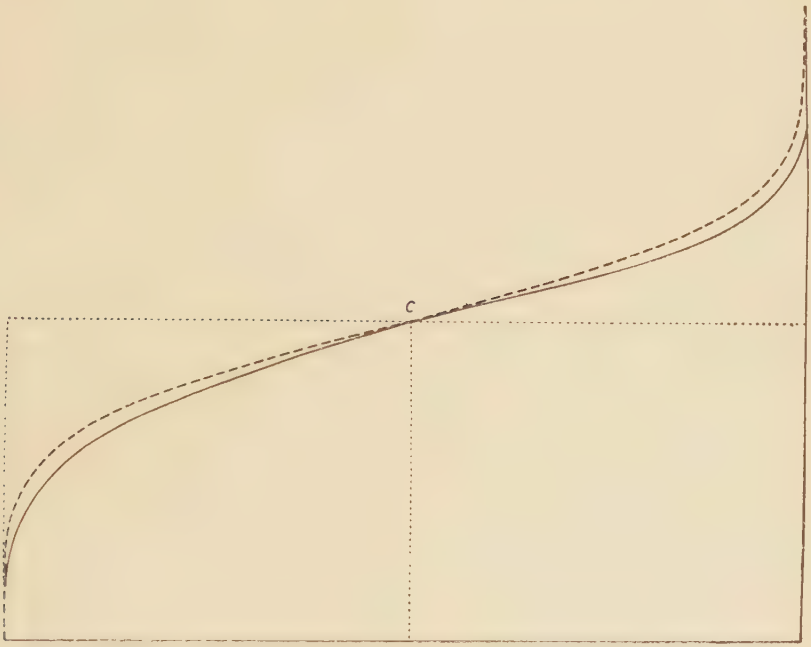


fig. 15

che tiene conto della dissomiglianza indipendentemente dal segno delle differenze, contrapporremo pertanto quello di *disnormalità* che tiene conto anche del segno di dette differenze.

In questo caso, vi è luogo a distinguere tra *disnormalità positiva*, *disnormalità nulla* e *disnormalità negativa* a seconda che le differenze positive superino, uguaglino o restino inferiori in valore assoluto alle differenze negative. S'intende che vi può essere disnormalità nulla quando vi sia anormalità; ma viceversa, quando una curva è normale, essa presenta sempre disnormalità nulla.

Adottando l'indice semplice di dissomiglianza, l'indice assoluto di disnormalità si riduce alla differenza tra la media arit-

metica della distribuzione osservata e la media aritmetica della corrispondente distribuzione normale, e l'indice relativo al rapporto di tale differenza al doppio dello scostamento semplice medio. Le rispettive formule sono (1):

$$^1N_p = A_x - A_y \tag{37.1}$$

$$^1n_p = \frac{A_x - A_y}{2\,{}^1S_M} \tag{38.1}$$

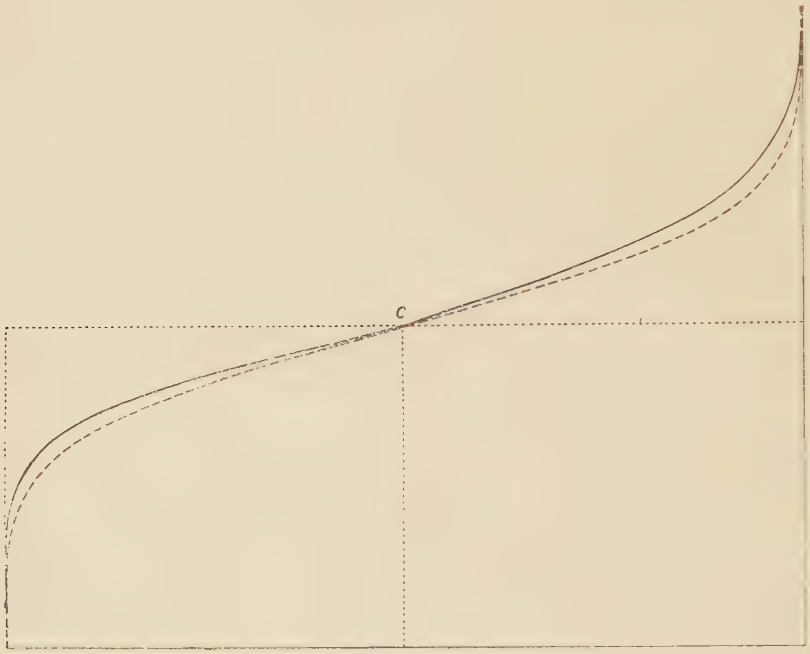


fig. 16

Esaminiamo le relazioni tra asimmetria e anormalità o disnormalità delle curve di graduazione.

Una curva normale è sempre simmetrica, ma una curva simmetrica può non essere normale, come è il caso per le curve continue rappresentate alle figg. 11, 12, 13, 14.

(1) Nell'edizione del 1947 del nostro *Corso di Statistica* e nella memoria inserita nell'« Enciclopedia delle Matematiche elementari » (cfr. nota 1 a p. 3) venne per errore omesso, nella formula (38.1), il fattore 2 che deve figurare al denominatore.

Una curva anormale può dunque essere simmetrica, come nelle figg. 11, 12, 13, 14, o asimmetrica, come nelle figg. 7, 8, 15 e 16.

Una curva anormale simmetrica ha sempre disnormalità nulla, ma una curva anormale asimmetrica può avere disnormalità positiva, nulla o negativa. Nella fig. 15 la curva anormale ha disnormalità negativa, mentre nella fig. 16 ha disnormalità positiva.

Quale è la relazione tra dissimetria e disnormalità delle curve?

Ci si persuade facilmente che, se una curva è dissimmetrica, è anche disnormale e, viceversa, se è disnormale, è anche dissimmetrica, e che, se ha dissimetria positiva, nulla o negativa, ha anche disnormalità positiva o rispettivamente nulla o negativa; non solo, ma che quanto più la dissimetria positiva o negativa è forte, tanto più è forte la disnormalità rispettivamente positiva o negativa. Converrebbe dunque avere, per i due aspetti della distribuzione, la stessa misura. Ed effettivamente le due misure da noi proposte praticamente coincidono. Basta infatti considerare che la media aritmetica della curva normale coincide con la sua mediana e che, per definizione, questa coincide con la mediana della curva osservata. Gli indici assoluto (37.1) e relativo (38.1) di disnormalità si riducono pertanto alle formule:

$$^1N_p = A_x - M_x \qquad (37.2)$$

$$^1n_p = \frac{A_x - M_x}{2\,{}^1S_M}\,, \qquad (38.2)$$

le quali praticamente coincidono con le formule già date per i rispettivi indici di dissimetria in quanto la (38.2) è uguale alla (25.2) a meno del fattore $\frac{1}{2}$ e la (37.2) è uguale alla (23.2) a meno del fattore $\frac{m}{n}$.

* * *

13. Passiamo ora all'esame della *curtosi* delle distribuzioni, e anzitutto fissiamo la terminologia.

La parola greca *curtosi* non richiama immediatamente alla mente del lettore il fenomeno che designa, tanto che si è dovuto ricorrere a dei giochi di parole per imprimere nella memoria che cosa significhino le espressioni *curve leptocurtiche* e *curve plati-*

curtiche (1). Le espressioni, pure usate, di *curve eccessive* o, viceversa, *curve deficienti* stanno a indicare la eccessività o, viceversa, la deficienza della curvatura, ma non indicano il termine di confronto rispetto a cui la curva viene giudicata eccessiva o deficiente e lo stesso inconveniente presentano le altre espressioni, a cui si è pensato o a cui si potrebbe pensare, di *curve appuntite* o *appiattite, aguzze* od *ottuse, sopraelevate* o *depresse*. Il termine di confronto è, invece, designato, insieme al concetto di eccessività, nelle espressioni, pure a volte impiegate, di *curve iperbinomiali* e *ipobinomiali*. In realtà, però, la curva effettiva non si confronta con la *curva binomiale*, così detta perchè le sue ordinate corrispondono ai termini di sviluppo del binomio $(p + q)^n$, ma con la *curva normale*, che è solo un caso particolare della curva binomiale per $p = q$ ed n infinitamente grande. Meglio quindi che di iperbinomialità ed ipobinomialità, si parlerà di *ipernormalità* ed *iponormalità* delle curve di distribuzione.

* * *

14. Come abbiamo detto, la ipernormalità o iponormalità è stata in origine considerata in relazione alle distribuzioni simmetriche, ed effettivamente, anche dal punto di vista logico, è bene considerare dapprima questo caso particolare.

Nel caso di distribuzioni simmetriche, chiamiamo, in tal senso ristretto, *ipernormale* (o viceversa *iponormale*) *una distribuzione la cui curva di frequenza presenta una maggiore* (*fig.* 10) (*o rispettivamente una minore – fig.* 9) *frequenza dei casi centrali ed estremi in confronto alla curva di frequenza normale, avente la stessa variabilità.*

La curva di graduazione corrispondente ad una distribuzione ipernormale simmetrica parte da un valore uguale (fig. 14) o inferiore (fig. 12) della rispettiva curva normale, decorre per un tratto sotto la curva normale e poi l'attraversa, mantenendosene al di sopra sino al valore mediano per seguire poi l'andamento opposto. Mentre la variabilità è, per definizione, uguale per le due curve, gli scostamenti estremi, corrispondenti alle ali della

(1) « Student » dà il seguente motto mnemonico: « Platykurtic curves, like the platypus, are squat with short tails. Leptokurtic curves are high with long tails, like the kangaroo — noted for « lepping »! Cfr. YULE and KENDALL, op. cit., pag. 165.

curva di graduazione, sono più alti e gli scostamenti intermedi, corrispondenti alla parte centrale della curva, sono più bassi nella distribuzione ipernormale che nella normale. Viceversa, nella distribuzione iponormale simmetrica, gli scostamenti estremi sono più bassi e gli intermedi più alti che nella normale; in essa, la curva di graduazione parte da un valore superiore (fig. 11) o uguale (fig. 13) a quello della curva normale, corre per un tratto sopra la curva normale e poi l'attraversa, mantenendosene al di sotto sino al valore mediano, per seguire poi l'andamento opposto.

Se noi moltiplichiamo l'intensità di ogni scostamento per la sua *distanza graduale* dalla mediana (misurata dal numero degli scostamenti intermedi + 1) [1] e facciamo la somma dei prodotti, otteniamo un totale che sarà, nella curva ipernormale, più alto e, nella curva iponormale, più basso che nella normale.

La differenza fra i due totali costituirà un punto naturale di partenza per la misura della ipernormalità o iponormalità della distribuzione.

Ora, è stato dimostrato che la somma dei prodotti degli scostamenti per la loro distanza graduale dalla mediana è uguale ad $\frac{1}{4}$ della somma delle $m(m-1)$ differenze che si possono istituire tra le m quantità della distribuzione, essendo

$$\Delta = \frac{4}{m(m-1)} \sum_{i=1}^{m} d_{i,\omega} \mid x_i - M \mid \qquad (39)$$

dove $d_{i,\omega}$ indica la distanza graduale della quantità x_i dalla mediana e Δ è la differenza media della serie [2].

D'altra parte, nella distribuzione normale è [3]

$$\Delta = \sqrt{2}\ {}^{1}S_M. \qquad (40)$$

L'indice assoluto di ipernormalità, o iponormalità, della distribuzione si potrà pertanto ricavare dalla quantità

---

(1) Per il concetto di *distanza graduale*, cfr. *Variabilità e Mutabilità*, 1ª edizione in « Studi economico-giuridici della R. Università di Cagliari », vol. 3°, parte 2ª, 1912, pag. 23; 2ª edizione in *Memorie di Metodologia Statistica*, Vol. I, *Variabilità e Concentrazione*, Milano, Giuffrè, 1939, pag. 209.

(2) Cfr., per la dimostrazione, *Variabilità e Mutabilità*, 2ª edizione, pag. 215.

(3) Cfr. *Variabilità e Mutabilità*, op. cit., 2ª ediz., pagg. 247-252.

$$Y = \Delta - \sqrt{2}\ {}^1S_M, \qquad (41)$$

che è positiva nel caso di ipernormalità e negativa in quello di iponormalità.

Per ottenere l'indice relativo, converrà ragguagliare Y al suo massimo positivo nel caso di ipernormalità, e al suo massimo negativo nel caso di iponormalità. Ora il valore di Δ non può mai salire sopra $2\,{}^1S_M$ nè scendere sotto ${}^1S_M$ (1). L'indice relativo di ipernormalità o di iponormalità si potrà pertanto desumere dalle formule

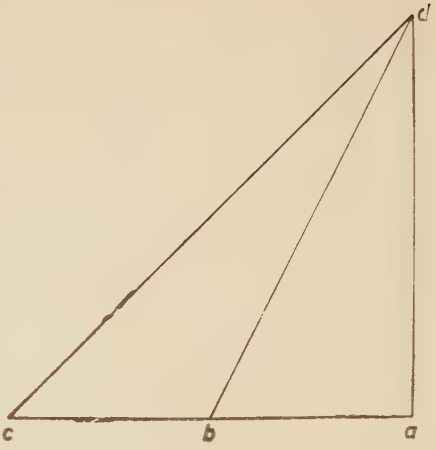


fig. 17

$$y' = \frac{\Delta - \sqrt{2}\ {}^1S_M}{2\,{}^1S_M - \sqrt{2}\ {}^1S_M} = \frac{\Delta}{0{,}5858\ {}^1S_M} - 2{,}4142 \qquad (42^1)$$

$$y'' = \frac{\Delta - \sqrt{2}\ {}^1S_M}{\sqrt{2}\ {}^1S_M - {}^1S_M} = \frac{\Delta}{0{,}4142\ {}^1S_M} - 3{,}4142 \qquad (42^2)$$

(1) Il rapporto tra «l'area di concentrazione» di una seriazione e l'area del triangolo, il quale ha come base la retta di equidistribuzione e come vertice opposto alla base il punto della curva di concentrazione corrispondente alla mediana, si sa essere uguale a $\frac{\Delta}{{}^1S_M}$

Ora questo rapporto è sempre compreso tra 1 e 2. E' precisamente

$$1 \leqq \frac{\Delta}{{}^1S_M} \leqq 2$$

Ora la prima disuguaglianza è evidente; della seconda è facile rendersi conto se si ricorda che la curva di concentrazione è concava verso l'asse delle y per tutto il suo andamento così che essa risulta interna al rettangolo che ha la stessa base e la stessa altezza del triangolo proporzionale ad ${}^1S_M$

L'uguaglianza $\frac{\Delta}{{}^1S_M} = 2$ ha luogo quando il carattere prende in $(n-1)$ casi uno stesso valore x_1 e in un solo caso un valore $x_2 > x_1$.

In particolare, se tutto l'ammontare del carattere è concentrato in un solo caso, essendo nullo negli altri $(n-1)$ casi, si ha ${}^1S_M = A$ e $\Delta = 2A$. Tracciando la curva di concentrazione (fig. 17), il valore di Δ risulta rappresentato in tale ipotesi dall'area del triangolo $c\ a\ d$, mentre il valore di ${}^1S_M$ risulta rappresentato dall'area del triangolo $c\ b\ d$.

di cui la prima, da usarsi quando sia $\Delta > \sqrt{2}\ {}^1S_M$, serve a misurare la ipernormalità e la seconda, da usarsi quando sia $\Delta < \sqrt{2}\ {}^1S_M$, serve a misurare la iponormalità.

* * *

15. Si avverta che le curve ipernormali o iponormali simmetriche, sopradefinite, costituiscono solo una categoria particolare di curve anormali simmetriche, particolarmente semplici. Le diremo curve *ipernormali o iponormali in senso stretto.* È ad esse che usualmente ci si riferisce nello studio della curtosi. Molte curve anormali simmetriche vi sono che non sono nè ipernormali nè iponormali nel senso stretto sopradefinito. È il caso, per esempio, delle curve 19 e 20.

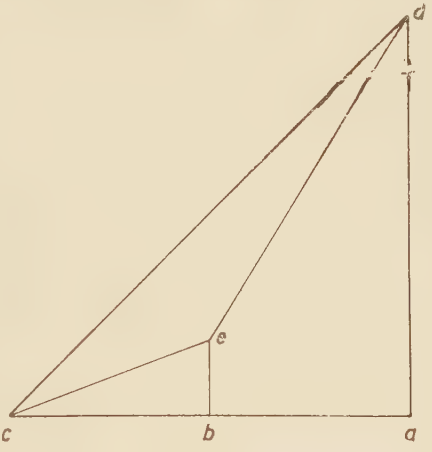


Fig. 18

Pur restando nel campo delle curve simmetriche, potremo estendere il concetto di ipernormalità e iponormalità, dicendo che *una curva anormale è ipernormale o iponormale in senso lato a seconda che, per ciò che ha riguardo alla relazione tra differenza media e scostamento semplice medio dalla mediana, si comporta come una curva ipernormale o, invece, iponormale in senso stretto, a seconda, cioè, che è*

$$\Delta \gtrless \sqrt{2}\ {}^1S_M\ .$$

Chiameremo *curva anormale simmetrica semplice* quella che, sia nel suo ramo destro che nel suo ramo sinistro, attraversa solo

---

Quando il carattere assume in una metà dei casi un'intensità e, nell'altra metà dei casi, un'altra intensità, è $\Delta_R = \frac{n-1}{n}\Delta = {}^1S_M$. Tracciando la curva di concentrazione (fig. 18), l'area del triangolo *c e d* rappresenta, in tale ipotesi, tanto il valore di Δ_R che il valore di ${}^1S_M$. Nel caso particolare in cui una metà dei casi abbia una intensità zero, i valori di Δ_R e ${}^1S_M$ risultano entrambi rappresentati dall'area del triangolo *c b d* (fig. 17).

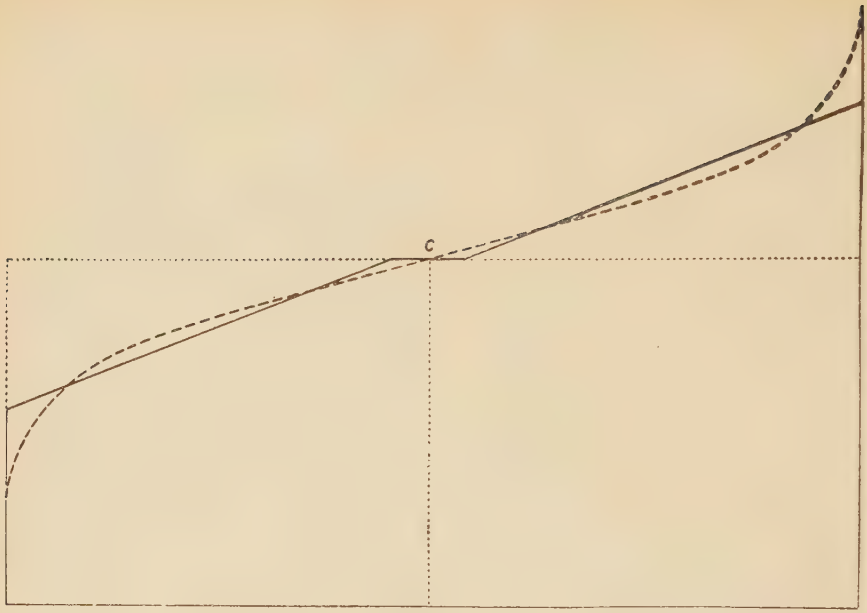


fig. 19

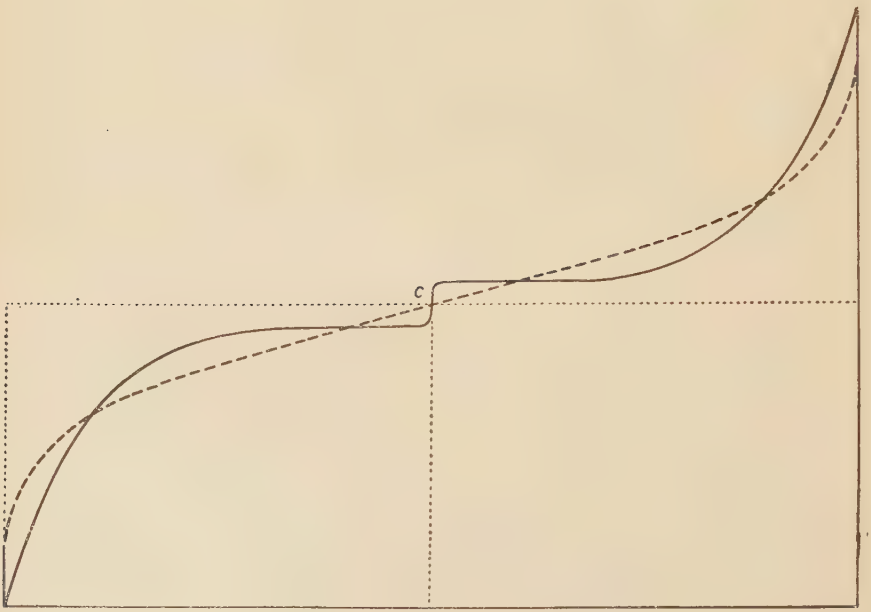


Fig. 20

una volta la corrispondente curva normale di graduazione, come avviene nelle figure 11, 12, 13, 14 e, invece, *curva anormale simmetrica complessa* quella che, in uno almeno dei due rami, attraversa la corrispondente curva di graduazione normale più di una volta, come nelle figure 19 e 20.

Le curve simmetriche ipernormali o iponormali in senso stretto, sono curve anormali simmetriche semplici e, viceversa, tutte le curve anormali simmetriche semplici sono necessariamente ipernormali o iponormali. Invece, le curve simmetriche che sono

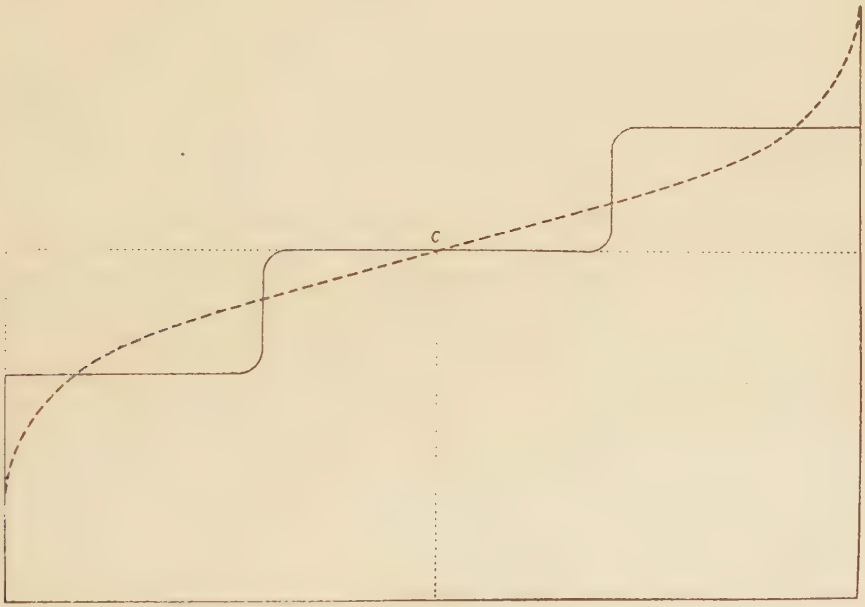


fig. 21

ipernormali o iponormali solo in senso lato sono curve anormali simmetriche complesse e, d'altra parte, le curve anormali simmetriche complesse possono essere ipernormali o iponormali, ma possono anche essere nè ipernormali nè iponormali, quando sia $\Delta = \sqrt{2}\,{}^1S_M$; esse si diranno allora *neutre*; tale è il caso della fig. 21.

Si potranno pertanto distinguere, tra le curve simmetriche anormali: *curve semplici ipernormali* (figure 12, 14), *curve semplici iponormali* (figure 11, 13), *curve complesse ipernormali* (fig. 20) *curve complesse iponormali* (fig. 19), *curve complesse neutre* (fig. 21).

Gli indici Y, y', y'' servono a misurare in tutti questi casi la ipernormalità, la iponormalità o la neutralità delle curve anormali.

* * *

16. D'altra parte, la questione della ipernormalità o iponormalità della distribuzione potrà evidentemente porsi anche per le distribuzioni asimmetriche, e nulla ci impedisce di generalizzare il concetto di ipernormalità o iponormalità e gli indici rela-

fig. 22

tivi, estendendo l'uno e gli altri anche al caso delle distribuzioni asimmetriche. La fig. 22 dà l'esempio di una distribuzione asimmetrica ipernormale e la fig. 23 di una distribuzione asimmetrica iponormale.

Per le curve anormali asimmetriche, non vale però la distinzione in semplici e complesse. Potrà darsi, infatti, che la curva di graduazione asimmetrica non attraversi la curva normale in nessun punto, come è il caso nelle figure 15 e 16. Potrà darsi pure che, in un ramo, la curva di graduazione anormale attraversi la curva normale una volta e, nell'altro ramo, nessuna o invece più di una volta, come è il caso nelle figure 24 e 25. Potrà darsi ancora che la

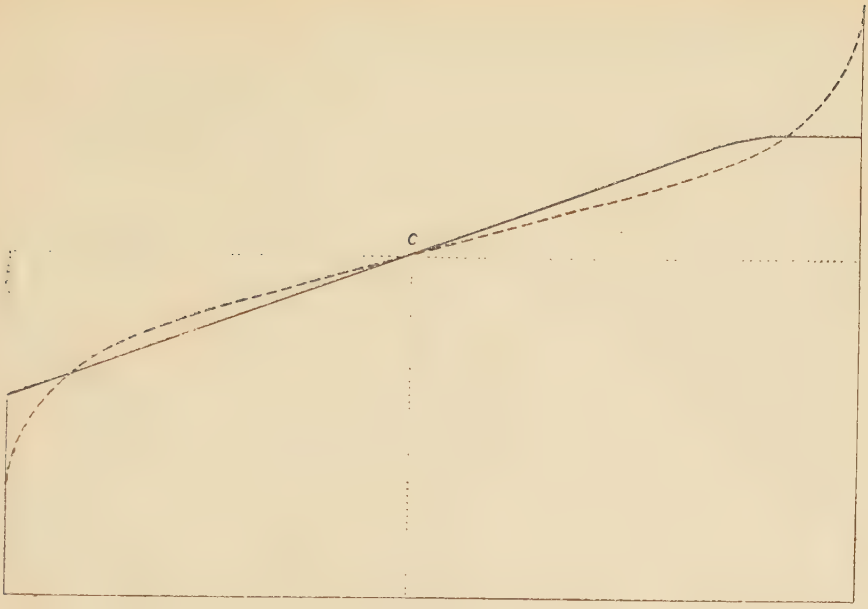


fig. 23

fig. 24

curva di graduazione osservata coincida, in un ramo, con la normale e, nell'altro, invece, ne differisca, come è il caso nelle figure 7 e 8. E potrà darsi, infine, che la curva di graduazione osservata, in un ramo, sia ipernormale e nell'altro, invece, iponormale, come nelle figure 27 e 28.

* * *

17. Osserviamo incidentalmente che, quando è stata da noi introdotta nella metodologia statistica la differenza media come

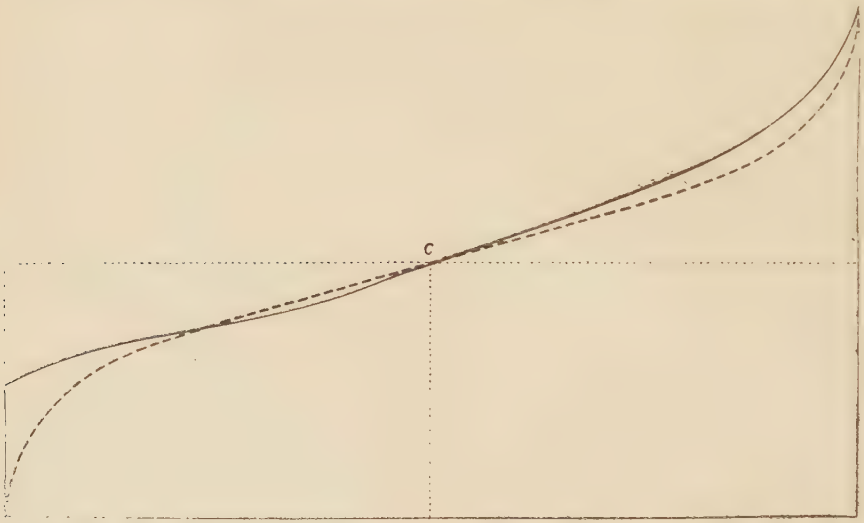


fig. 25

indice di variabilità, è stato osservato da qualcuno che tale indice era inutile perchè non dava risultati diversi da quelli ottenuti con lo scostamento semplice medio ed anzi rappresentava, di fronte a quest'ultimo, uno svantaggio, in quanto di meno facile calcolo; e la critica — per quanto subito dimostrata infondata — fu ripetuta poi, provocando altre rettifiche(1).

Veramente, prima ancora che fosse pubblicata la memoria *Variabilità e Mutabilità*, in cui era proposta la differenza media,

(1) Cfr., per i vari autori che si sono occupati di tale questione, *Variabilità e Mutabilità*, 2ª edizione, nota 2 a pagg. 279-281 e G. Pietra, *I contributi degli Italiani al progresso della Statistica metodologica*, nell'opera: *Un secolo di progresso scientifico italiano*, Società Italiana per il Progresso delle Scienze, 1939, pagg. 295-297.

noi avevamo avuto cura di far riscontrare se l'adozione di detta differenza media, come indice di variabilità, poteva portare a conclusioni diverse dall'adozione dello scostamento semplice medio o dello scostamento quadratico medio, e le numerose applicazioni eseguite avevano posto fuori dubbio tale possibilità(1), confermata poi da ricerche ulteriori (2).

Si sarebbe potuto — è vero — domandarsi se si trattasse veramente di differenze sistematiche o non piuttosto di differenze accidentali ; ma anche questo dubbio era stato eliminato dopo che venne dimostrato come la divergenza tra differenza media e scostamento semplice medio dipende dalla forma della curva di concentrazione e può quindi essere sistematica (3).

Le osservazioni che precedono, non solo confermano il carattere sistematico della divergenza, ma mostrano come questa dipenda precisamente dalla ipernormalità o iponormalità della distribuzione e possa essere pertanto opportunamente assunta a base della misura di tale ipernormalità o iponormalità, come noi appunto abbiamo fatto in questo articolo.

* * *

18. Nel numero 16, abbiamo accennato alla possibilità che una curva di graduazione sia ipernormale o iponormale, in un ramo, e normale, nell'altro, oppure ipernormale in un ramo e iponormale nell'altro. Questa constatazione suggerisce l'opportunità di studiare separatamente l'anormalità dei due rami di curve.

Potrà darsi che la curva sia anormale solo in un ramo o, invece, in entrambi ; distingueremo pertanto una *anormalità unilaterale* da una *anormalità bilaterale*.

Quando vi è anormalità unilaterale, vi potrà essere *ipernormalità unilaterale*, *iponormalità unilaterale* o *anormalità neutra unilaterale*. Quando vi è anormalità bilaterale, vi potrà essere, in

---

(1) Cfr. G. Dettori, *Contributi allo studio della variabilità dei prezzi*, in « Studi economico-giuridici della R. Università di Cagliari », Vol. IV, parte 1ª, 1912.

(2) Cfr. in proposito *Variabilità e Mutabilità*, 2ª edizione, pagg. 278-279 e le opere ivi citate alla nota (1) di pag. 279.

(3) Cfr. G. Pietra, *Delle relazioni tra gli indici di variabilità*, nota I e nota II, in « Atti del R. Istituto Veneto di Scienze, Lettere ed Arti », 1914-15, Tomo LXXIX, parte II.

entrambi i rami, ipernormalità o iponormalità o anormalità neutra, nel qual caso potremo parlare di *ipernormalità uniforme* o rispettivamente di *iponormalità uniforme* o di *anormalità neutra uniforme;* e vi potrà essere, in un ramo, ipernormalità e nell'altro iponormalità, oppure, in un ramo, anormalità neutra e, nell'altro, ipernormalità o iponormalità, nel qual caso, secondo che prevale la ipernormalità o la iponormalità o non prevale nessuna delle due, diremo che vi è *ipernormalità disforme* o *iponormalità disforme* o *anormalità neutra disforme.*

Una distinzione corrispondente non si presenta invece per la simmetria e per la dissimetria, in quanto ogni loro indice già coinvolge tanto il ramo destro che il ramo sinistro della curva, tra i quali, appunto, avviene il confronto. Nè si presenta per la disnormalità, in quanto questa è sempre bilaterale e uguale per i due rami, per il fatto che la somma degli scostamenti della distribuzione disnormale è, per ipotesi, uguale alla somma degli scostamenti della distribuzione normale con cui si confronta, per modo che, se in uno dei rami prevalgono le differenze positive a favore della distribuzione normale, nell'altro ramo devono di altrettanto prevalere le differenze positive a favore della distribuzione disnormale.

Gli indici di anormalità e di ipernormalità (o iponormalità) sovraesposti valgono evidentemente qualunque sia la forma della curva considerata, sia questa anormale unilaterale o bilaterale ; o, se bilaterale, in modo uniforme o disforme nei due rami. Si tratta di *indici globali di anormalità* o *di ipernormalità* (o *iponormalità*).

Lo studio della conformazione delle curve può però spingersi più oltre in quanto si passi a considerare l'anormalità e, in particolare, la ipernormalità distintamente per i due rami della curva, il destro e il sinistro.

Con gli stessi criteri con cui si sono stabiliti indici globali, si potranno stabilire *indici parziali di anormalità* e *di ipernormalità* (o *iponormalità*) distintamente per il ramo destro e per il ramo sinistro della curva.

Posto, al solito, $n = \frac{m}{2}$ per m pari, $n = \frac{m+1}{2}$ per m dispari, l'*indice assoluto destro di anormalità* sarà

$$ {}^1N_d = \frac{1}{m} \sum_{i=n+1}^{m} | x_i - y_i | \qquad (43.1) $$

e l'*indice assoluto sinistro di anormalità* sarà

$$ {}^1N_s = \frac{1}{m} \sum_{i=1}^{n} | x_i - y_i | \quad . \tag{43.2} $$

che, ragguagliati al massimo (il quale è, per entrambi, $= {}^1S_M$) daranno luogo all'*indice relativo destro di anormalità*

$$ n_d = \frac{{}^1N_d}{{}^1S_M} = \frac{\sum_{i=n+1}^{m} | x_i - y_i |}{\sum_{i=n+1}^{m} (y_i - M)} \tag{44.1} $$

e rispettivamente all'*indice relativo sinistro di anormalità*

$$ {}^1n_s = \frac{{}^1N_s}{{}^1S_M} = \frac{\sum_{i=1}^{n} | x_i - y_i |}{\sum_{i=1}^{n} (M - y_i)} \quad . \tag{44.2} $$

Similmente, l'*indice assoluto destro di ipernormalità* (o *iponormalità*) sarà fornito dalla formula:

$$ Y_d = \Delta_d - \frac{1}{\sqrt{2}} {}^1S_M \tag{45.1} $$

e l'*indice assoluto sinistro di ipernormalità* (o *iponormalità*) dalla formula

$$ Y_s = \Delta_s - \frac{1}{\sqrt{2}} {}^1S_M \ , \tag{45.2} $$

dove è

$$ \Delta_d = \frac{4}{m(m-1)} \sum_{i=n+1}^{m} d_{i,M} (x_i - M) $$

$$ \Delta_s = \frac{4}{m(m-1)} \sum_{i=1}^{n} d_{i,M} (M - x_i) \ . $$

Ragguagliando, al solito, gli indici parziali assoluti di ipernormalità al rispettivo massimo, si ottengono gli indici parziali

relativi di ipernormalità. L'*indice relativo destro di ipernormalità* o *iponormalità* si ricaverà dalla formula

$$y_d' = \frac{\Delta_d}{0{,}2929\ {}^1S_M} - 2{,}4142 \qquad (46^1.1)$$

o rispettivamente dalla

$$y_d'' = \frac{\Delta_d}{0{,}2071\ {}^1S_M} - 3{,}4142, \qquad (46^2.1)$$

di cui si dovrà usare la prima o la seconda, a seconda che sia

$$Y_d \gtrless 0$$

e, analogamente, l'*indice relativo sinistro di ipernormalità* o *iponormalità* dalla formula

$$y_s' = \frac{\Delta_s}{0{,}2929\ {}^1S_M} - 2{,}4142 \qquad (46^1.2)$$

o rispettivamente dalla

$$y_s'' = \frac{\Delta_s}{0{,}2071\ {}^1S_M} - 3{,}4142, \qquad (46^2.2)$$

di cui si dovrà usare corrispondentemente la prima o la seconda a seconda che sia

$$Y_s \gtrless 0\,.$$

* * *

19. Esaminiamo le relazioni che intercedono tra indici globali e indici parziali di anormalità o di ipernormalità (o iponormalità). Si ha

$${}^1N_d + {}^1N_s = {}^1N$$

$$Y_d + Y_s = Y$$

uguaglianze che mostrano come la *somma dei due indici parziali assoluti di anormalità, o di ipernormalità* (o *iponormatità*) *è uguale*

all'indice globale assoluto di anormalità, o rispettivamente di ipernormalità, (o *iponormalità*) *della stessa distribuzione.*

Si ha pure

$$\frac{{}^1n_d + {}^1n_s}{2} = {}^1n$$

che mostra come *l'indice globale relativo d'anormalità è uguale alla media aritmetica semplice degli indici parziali relativi di anormalità della stessa distribuzione.* Si hanno infine le uguaglianze

$$\frac{y_d' + y_s'}{2} = y'$$

$$\frac{y_d'' + y_s''}{2} = y''$$

che mostrano come, nel caso di ipernormalità o iponormalità bilaterale uniforme, l'indice globale di ipernormalità o iponormalità è uguale alla media aritmetica semplice degli indici parziali rispettivi. Nel caso, invece, di ipernormalità o iponormalità disforme, l'indice globale di ipernormalità o iponormalità sta, con gli indici parziali rispettivi, nella relazione

$$\frac{0{,}2929\, y_d' + 0{,}2071\, y_s''}{0{,}5858} = y'$$

oppure

$$\frac{0{,}2071\, y_d'' + 0{,}2929\, y_s'}{0{,}5858} = y'$$

nel caso di ipernormalità; e invece nella relazione

$$\frac{0{,}2929\, y_d' + 0{,}2071\, y_s''}{0{,}4142} = y''$$

oppure

$$\frac{0{,}2071\, y_d'' + 0{,}2929\, y_s'}{0{,}4142} = y''$$

nel caso di iponormalità.

* * *

20. Misurate così l'anormalità o l'ipernormalità dei due rami della curva, noi potremo porci l'ulteriore questione se ed in quale dei due rami l'anormalità o l'ipernormalità della distribuzione sia più accentuata.

Per ciò che concerne l'anormalità, le espressioni

$$P_n = N_d - N_s \tag{47}$$

$$p_n = n_d - n_s \tag{48}$$

ci forniscono l'*indice assoluto*, o rispettivamente l'*indice relativo, di prevalenza dell'anormalità.*

Diremo che vi è *prevalenza destra* (figg. 7 e 22) oppure *prevalenza sinistra* (fig. 8) *dell'anormalità* a seconda che l'indice risulti maggiore o minore di zero.

Nel caso di anormalità bilaterale, potrà anche darsi che sia $N_d = N_s$ e quindi $n_d = n_s$, nel qual caso diremo che vi è *equivalenza dell'anormalità.*

In modo analogo, potremo stabilire gli *indici assoluti e relativi di prevalenza della ipernormalità* o *iponormalità*, desumendoli dalle formule

$$P_y = |Y_d| - |Y_s| \tag{49}$$

$$p_y = |y_d| - |y_s| \tag{50}$$

Essi possono assumere valori positivi, nulli o negativi a seconda che vi sia prevalenza destra, equivalenza o prevalenza sinistra dell'ipernormalità o iponormalità. Qui le differenze si istituiscono tra valori assoluti degli indici di ipernormalità o iponormalità, in quanto questi possono assumere valori positivi o negativi. I risultati possibili saranno quindi molto più numerosi.

Vi potrà essere, invero, *ipernormalità unilaterale destra* (fig. 7) o *sinistra* e, similmente, *iponormalità unilaterale* pure *destra* o invece *sinistra* (fig. 8).

Vi potrà essere *ipernormalità bilaterale uniforme con prevalenza destra* (fig. 22), *con equivalenza* (fig. 12 e 14) o *con prevalenza sinistra* e, similmente, *iponormalità bilaterale uniforme con preva-*

lenza destra (fig. 23), o *con equivalenza* (figg. 11 e 13), o *con prevalenza sinistra.*

Vi potrà, infine, essere *ipernormalità disforme con prevalenza destra* (fig. 26) o viceversa *sinistra* e, similmente, *iponormalità disforme con prevalenza destra* o viceversa *sinistra* (fig. 27).

fig. 26

Quando poi, in uno dei rami della curva, vi sia ipernormalità e, nell'altro ramo, iponormalità della stessa intensità, diremo che vi è *equivalenza tra ipernormalità e iponormalità* quando l'ipernormalità è sul ramo destro (fig. 28), e *equivalenza tra iponormalità e ipernormalità* quando l'ipernormalità è, al contrario, sul ramo sinistro.

Quando vi sia ipernormalità o iponormalità unilaterale, uno degli indici parziali Y_d e Y_s è nullo, *e quando vi sia ipernormalità o iponormalità disforme,* i due indici parziali hanno segni diversi. In entrambi i casi, è

$$P_y = |\, Y_d \,| - |\, Y_s \,| = |\, Y_d + Y_s \,| = |\, Y \,|\,; \qquad (49\ bis)$$

vale a dire, l'*indice assoluto di prevalenza dell'ipernormalità è, in valore assoluto, uguale all'indice assoluto di ipernormalità.*

Quando vi sia, invece, ipernormalità o iponormalità bilaterale uniforme, il primo è minore del secondo avendosi

$$P_y = |Y_d| - |Y_s| < |Y_d + Y_s| = |Y|$$

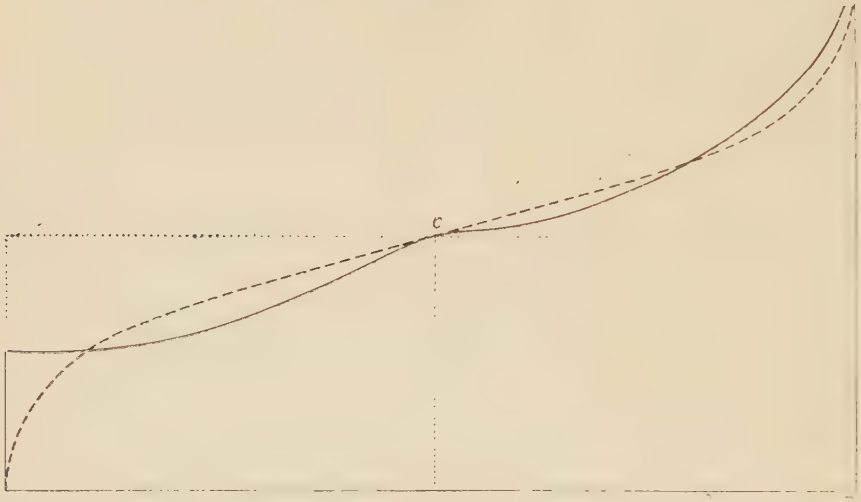


fig. 27

e sarà

$$P_y = |Y_d| - |Y_s| = |\Delta_d - \Delta_s|\,; \qquad (50\ ter)$$

Perciò che riguarda gli indici relativi, quando vi sia ipernormalità o iponormalità unilaterale oppure ipernormalità o iponormalità disforme, sarà

$$p_y = \frac{0{,}2071\,\Delta_d + 0{,}2929\,\Delta_s}{0{,}2929.0{,}2071\,{}^1S_M} - 5{,}8283, \qquad (50^1\ bis)$$

nel caso di ipernormalità nel ramo destro;

$$p_z = 5{,}8283 - \frac{0{,}2929\,\Delta_d + 0{,}2071\,\Delta_s}{0{,}2929\,.\,02071\,{}^1S_M}\,, \qquad (50^2\ bis)$$

nel caso di ipernormalità nel ramo sinistro. Quando vi sia invece ipernormalità o iponormalità bilaterale uniforme, sarà

$$p_y = \frac{\Delta_d - \Delta_s}{0{,}2929\,{}^1S_M} \qquad (50^1\ ter)$$

nel caso di ipernormalità bilaterale uniforme, e

$$p_y = \frac{\Delta_d - \Delta_s}{0{,}2071\ {}^1S_M} \qquad (50^2\ ter)$$

nel caso di iponormalità bilaterale uniforme.

* * *

21. Vi sono curve che, mentre sono nettamente anormali, si possono considerare come risultanti di due rami di curve normali dotate di differente variabilità. Tale è il caso, per esempio,

fig. 28

nelle figg. 15 e 16. L'esistenza di curve siffatte fa sorgere l'opportunità della determinazione di indici alquanto diversi dai precedenti, che chiameremo *indici laterali* e che hanno per iscopo di confrontare la conformazione del ramo destro (o sinistro) della curva con una semicurva normale, avente la stessa variabilità.

Si intende facilmente che gli indici laterali vengono a coincidere con gli *indici parziali* sopraconsiderati, quando sia $A = M$ e quindi ${}^1S_d = {}^1S_s$, dove è

$${}^1S_d = \frac{1}{n} \sum_{i=n+1}^{m} (x_i - M)$$

$${}^1S_s = \frac{1}{n} \sum_{i=1}^{m} (M - x_i)\ ;$$

ma, quando invece sia ${}^1S_d \lesseqgtr {}^1S_s$, i due indici possono differire notevolmente.

Se con x_i si indica al solito il termine generico i^{mo} della distribuzione osservata, con $y_{s,i}$ $(i = 1, 2, \ldots, n)$ il termine i^{mo} della semidistribuzione normale di valori inferiori alla mediana, con $y_{d,i}$ $(i = n+1, n+2, \ldots, m)$ il termine i^{mo} della semidistribuzione normale di valori superiori alla mediana, l'*indice laterale destro assoluto di anormalità* sarà

$$ {}^1N_{ld} = \frac{1}{n} \sum_{i=n+1}^{m} | x_i - y_{d,i} | \qquad (51.1) $$

e l'*indice laterale destro relativo di anormalità*

$$ {}^1n_{ld} = \frac{{}^1N_{ld}}{{}^1S_d} = \frac{\sum_{i=n+1}^{m} | x_i - y_{d,i} |}{\sum_{i=n+1}^{m} (x_i - M)} \qquad (52.1) $$

e analogamente l'*indice laterale sinistro assoluto di anormalità* sarà

$$ {}^1N_{ls} = \frac{1}{n} \sum_{i=1}^{n} | x_i - y_{s,i} | \qquad (51.2) $$

e l'*indice laterale sinistro relativo di anormalità*

$$ {}^1n_{ls} = \frac{{}^1N_{ls}}{{}^1S_s} = \frac{\sum_{i=1}^{n} | x_i - y_{s,i} |}{\sum_{i=1}^{n} (M - x_i)} \qquad (52.2) $$

L'*indice laterale destro assoluto di ipernormalità* (o *iponormalità*) sarà

$$ Y_{ld} = \Delta_d - \frac{1}{\sqrt{2}} {}^1S_d \qquad (53) $$

e l'*indice laterale destro relativo di ipernormalità* (o *iponormalità*) sarà

$$ y'_{ld} = \frac{\Delta_d}{0{,}2929\ {}^1S_d} - 2{,}4142 \qquad (54^1) $$

e rispettivamente

$$ y''_{ld} = \frac{\Delta_d}{0{,}2071\ {}^1S_d} - 3{,}4142 \qquad (54^2) $$

dove sarà da usarsi la prima o la seconda formula a seconda che sia $Y_{ld} \gtrless 0$. Analoghe sono le formule per l'*indice assoluto o relativo bilaterale sinistro di ipernormalità o iponormalità.*

La prevalenza della anormalità o della ipernormalità laterale destra o sinistra si può misurare con formule analoghe a quelle usate per gli indici parziali.

* * *

22. Lo svolgimento della teoria degli indici di anormalità fatto nelle pagine precedenti è stato parallelo a quello degli indici di asimmetria. Il parallelismo fa però difetto in un punto. Mentre, per gli indici di asimmetria, noi abbiamo considerato due massimi — un massimo assoluto, vincolato alla conservazione dell'ammontare totale del carattere in esame, ed un massimo relativo, vincolato alla conservazione della distribuzione del carattere — per gli indici di anormalità, abbiamo considerato solo il massimo sottoposto a questo secondo vincolo. Ciò è dipeso dal fatto che una curva normale, con cui la distribuzione effettiva si confronta, non si può dire in generale definita se non si conosce un valore medio e, ad un tempo, un indice di variabilità. Senonchè, se questo è vero in generale, non è vero nel caso particolare delle distribuzioni di grandezze intensive, che esprimono la frequenza di un carattere, perchè allora il valore medio e gli indici di variabilità della curva normale sono legati da una relazione costante, cosicchè basta conoscere l'uno di essi perchè la curva normale sia definita. Infatti, se p è la probabilità del carattere (o, ciò che praticamente può considerarsi equivalente, la sua frequenza media (aritmetica) in un numero molto grande di osservazioni), lo scostamento quadratico medio dalla media aritmetica della distribuzione delle frequenze osservate in gruppi di s osservazioni — supposto p costante nelle singole osservazioni — ha per valore probabile $\sqrt{\frac{p\,(1-p)}{s}}$ e lo scostamento semplice medio dalla media aritmetica ha per valore probabile $\sqrt{\frac{2\,p\,(1-p)}{\pi\,s}}$. In tal caso, vi ha dunque modo di costruire anche indici di anormalità ragguagliati al loro massimo assoluto.

Si potrà osservare che, a rigore, una distribuzione normale di frequenza suppone che la frequenza media aritmetica (che

in questo caso coincide con la frequenza mediana) sia $= \frac{1}{2}$ e, se si dovesse limitarsi a determinare gli indici di anormalità ragguagliati al loro massimo assoluto quando la frequenza media aritmetica risulta, sia pure approssimativamente, $= \frac{1}{2}$, il loro campo di applicazione sarebbe molto ristretto.

Per superare questa difficoltà, si potrebbe pensare a generalizzare gli *indici di anormalità* in *indici di abinomialità*, diretti a paragonare la distribuzione delle frequenze osservate in certi gruppi di osservazioni con la distribuzione binomiale delle corrispondenti probabilità ; senonchè la determinazione di detta distribuzione binomiale, quando sono numerose le osservazioni dei singoli gruppi, è molto laboriosa. Soccorre però la considerazione che, quando le osservazioni sono numerose, la distribuzione binomiale, per valori di p non troppo diversi da $\frac{1}{2}$, si avvicina sufficientemente, agli scopi pratici, alla distribuzione normale, in quanto gli scostamenti estremi, in cui si manifesta l'asimmetria della distribuzione, sono così rari che in pratica dalla loro asimmetria si può prescindere. Non meno importante è il fatto che, nella gran parte dei casi, la variabilità osservata è nettamente più forte della variabilità teorica, cosicchè le rispettive curve di graduazione non si intersecano che in corrispondenza alla mediana, come avviene nella fig. 29. In tal caso, gli indici semplici o quadratici di dissomiglianza dipendono dalle differenze tra gli scostamenti semplici medi o tra gli scostamenti quadratici medi delle due curve, mentre la forma di queste viene a perdere ogni importanza.

L'indice di anormalità si riduce allora ad un *indice di dispersione* [1], il quale ha al numeratore la differenza tra gli scostamenti

---

[1] La parola «dispersione» significa, in origine, distribuzione dei valori, ma da questo significato originario sono derivati due significati leggermente diversi: secondo l'uno, con dispersione si vuol significare la intensità degli scostamenti rispetto alla media e in tal senso viene usata sopratutto dagli autori anglosassoni, per i quali, quindi, è sinonimo della parola «variabilità» usata dagli Italiani. In un altro senso, dispersione significa invece la distribuzione degli scostamenti paragonata alla distribuzione teorica che si verificherebbe per puro effetto del caso. È in tal senso che si parla della «teoria della dispersione» e di «indice di dispersione». L'indice di dispersione viene di solito desunto dal rapporto di una media degli scostamenti empirici al valore teorico che tale media assumerebbe per puro effetto del caso. Il Dormoy, che per primo propose tale indice (1874)

semplici medi (della distribuzione osservata e della teorica) dalla mediana, se si adotta l'indice semplice di dissomiglianza, o la radice quadrata della differenza tra i quadrati degli scostamenti quadratici medi dalla mediana, se si adotta l'indice quadratico

fig. 29

di dissomiglianza, e al denominatore il valore massimo che il rispettivo numeratore potrà assumere, ferma restando la frequenza media p. Potremo parlare rispettivamente di *indice sem-*

---

sotto il nome di «coefficiente di divergenza», ragguagliava lo scostamento semplice medio empirico allo scostamento semplice medio teorico. Il Lexis, che, poco dopo (1876), trattò più compiutamente l'argomento, ragguagliava lo scostamento quadratico medio empirico allo scostamento quadratico medio teorico. Usualmente si dice che la dispersione è *normale*, quando il rapporto è uguale ad 1; *supernormale*, quando è superiore ad 1 e *subnormale*, quando è inferiore ad 1. Normale, in questo senso, non significa conforme alla curva gaussiana, secondo il significato che si dà alla parola nell'espressione «curva normale», ma significa conforme alla curva binomiale che può differire sensibilmente, quando p sia molto diverso da $\frac{1}{2}$, da una curva normale; perciò ritengo preferibile di parlare non di dispersione normale, supernormale o subnormale, ma di dispersione *binomiale*, *iperbinomiale* o *ipobinomiale*.

plice e di *indice quadratico di dispersione*. Le formule saranno, per l'indice semplice di dispersione,

$$ {}^1Z' = \frac{{}^1S_M - {}^1s_M}{Max\,({}^1S_M) - {}^1s_M} \qquad (55^1) $$

(dove ${}^1s_M$ indica il valore dello scostamento semplice medio dalla mediana della distribuzione teorica ed è

$$ Max\,({}^1S_M) = \begin{cases} p \\ 1 - p \end{cases} $$

a seconda che sia $p \lesseqgtr 1/2$); e, per l'indice quadratico di dispersione,

$$ {}^2Z' = \sqrt{\frac{{}^2S^2_M - {}^2s^2_M}{Max\,({}^2S^2_M) - {}^2s^2_M}} \qquad (56^1) $$

(dove ${}^2S_M$, ${}^2s_M$ indicano gli scostamenti quadratici medi dalla mediana della distribuzione osservata e, rispettivamente, della teorica ed è

$$ Max\,({}^2S^2_M) = \begin{cases} p \\ 1 - p \end{cases} $$

a seconda che sia $p \lesseqgtr 1/2$).

Queste formule sono da adottarsi nel caso, che è di gran lunga il più frequente, in cui il numeratore sia > 0 (*dispersione iperbinomiale*). Nel caso che il numeratore fosse invece < 0 (*dispersione ipobinomiale*), dovremmo ricorrere alle formule seguenti:

$$ {}^1Z'' = \frac{{}^1s_M - {}^1S_M}{{}^1s_M} \qquad (55^2) $$

$$ {}^2Z'' = \frac{\sqrt{{}^2s_M{}^2 - {}^2S_M{}^2}}{{}^2s_M} \qquad (56^2) $$

L'applicazione di questi indici incontra però la difficoltà della determinazione dei valori di ${}^1s_M$ ed ${}^2s_M$ per i quali non si conoscono espressioni generali, valide per qualunque valore di p.

La difficoltà si supera se l'indice di variabilità si desume, anzichè dallo scostamento semplice o quadratico medio dalla mediana, dallo scostamento quadratico medio dalla media aritmetica, in quanto la sua espressione

$$ s^2_A = \sqrt{\frac{p\,(1-p)}{s}} \qquad (57) $$

ha, com'è noto, validità generale, qualunque sia il valore di p.

L'indice quadratico di *dispersione iperbinomiale* diviene allora

$$ {}^2Z' = \sqrt{\frac{{}^2S_A{}^2 - \frac{p\,(1-p)}{s}}{Max\,({}^2S_A)^2 - \frac{p\,(1-p)}{s}}} \qquad (58^1.1) $$

o anche, tenendo presente che è $Max\,({}^2S_A) = \sqrt{p\,(1-p)}$ (1)

$$ {}^2Z' = \sqrt{\frac{{}^2S_A{}^2 - \frac{p\,(1-p)}{s}}{p\,(1-p)\,\frac{s-1}{s}}} \qquad (58^1.2) $$

dove il fattore $\frac{s-1}{s}$ si può trascurare per s abbastanza grande.

Il numeratore delle formule (58^1) è un'espressione ben nota, già proposta dal Lexis per misurare la *componente fisica della dispersione* (2) vale a dire, lo scostamento quadratico medio che si verificherebbe quando fossero eliminate le differenze tra le frequenze osservate aventi carattere accidentale e rimanessero pertanto solo le differenze aventi carattere sistematico, quali si verificherebbero se il numero delle osservazioni fosse per ogni termine infinitamente grande. L'indice quadratico di dispersione ricavato dalle (58^1), costituisce dunque il rapporto della componente fisica della dispersione, secondo la formula del Lexis, al suo massimo.

(1) Il valore di ${}^2S_A$ sarà massimo quando in tutti gli m termini, la frequenza osservata sia $= 1$ oppure $= 0$. Tutti gli m termini avendo lo stesso numero di osservazioni uguale ad s, se p è la frequenza media del carattere nelle ms osservazioni, la sua frequenza sarà uguale ad 1 in pm termini, e uguale a 0 in $(1-p)\,m$ termini e si avrà :

$$ Max\,({}^2S_A) = \sqrt{\frac{1}{m}\,[mp\,(1-p)^2 + m\,(1-p)\,p^2]} = \sqrt{p\,(1-p)} $$

(2) Cfr. W. Lexis, *Abhandlungen zur Theorie der Bevölkerungs- und Moralstatistik*, Jena, Fischer, 1903, pag. 177 e segg.

Nel caso di *dispersione ipobinomiale*, l'indice sarà invece ricavato dalla formula

$$ {}^2Z'' = \sqrt{\frac{\frac{p\,(1-p)}{s} - {}^2S^2{}_A}{\frac{p\,(1-p)}{s}}} \qquad (58^2) $$

23. La circostanza che di solito la probabilità p del carattere non è conosciuta esattamente fa sì che essa sia in pratica sostituita, nelle (58), dalla sua frequenza nelle ms osservazioni, nel qual caso le (58) danno solo un valore approssimato dell'indice quadratico di dispersione; ma è facile dare anche la formula per il valore esatto di detto indice quadratico e mostrare che questo in pratica differisce da quello fornito dalle (58) per fattori trascurabili.

Si dimostra infatti che in tal caso il valore teorico di ${}^2S_A$ è dato esattamente, anzichè dalla (57), dalla

$$ {}^2s_A = \sqrt{p\,(1-p)\,\frac{m-1}{ms-1}} = \sqrt{\frac{p\,(1-p)}{s}\,\frac{ms-s}{ms-1}} \qquad (59) $$

e in generale, in luogo della ($58^1.2$) e (58^2), si ottengono le espressioni seguenti:

$$ {}^2Z' = \sqrt{\frac{{}^2S_A^2 - p\,(1-p)\,\frac{m-1}{ms-1}}{p\,(1-p)\,\frac{m\,(s-1)}{ms-1}}} = \sqrt{\frac{{}^2S_A^2 - \frac{p\,(1-p)}{s}\,\frac{ms-s}{ms-1}}{p\,(1-p)\,\frac{s-1}{s}\,\frac{ms}{ms-1}}} \qquad (60^1.1) $$

$$ {}^2Z'' = \sqrt{\frac{p\,(1-p)\,\frac{m-1}{ms-1} - {}^2S^2{}_A}{p\,(1-p)\,\frac{m-1}{ms-1}}} = \sqrt{\frac{\frac{p\,(1-p)}{s}\cdot\frac{ms-s}{ms-1} - {}^2S^2{}_A}{\frac{p\,(1-p)}{s}\cdot\frac{ms-s}{ms-1}}} \qquad (60^2) $$

Le formule (59) e (60) espresse nella seconda forma mostrano come esse differiscono dalle rispettive (57) e (58) per i coefficienti

$\frac{ms-s}{ms-1}$, $\frac{ms}{ms-1}$ che in pratica si possono considerare uguali ad 1 e quindi trascurare.

È anche da tener presente che la componente fisica della dispersione non è data esattamente, come il Lexis riteneva, dalla espressione:

$$R = \sqrt{{}^2S^2_A - \frac{p\,(1-p)}{s}}$$

ma dall'espressione:

$$R' = \sqrt{\frac{s}{s-1}\left[{}^2S^2_A - \frac{p\,(1-p)}{s}\right]}$$

se la probabilità p è conosciuta esattamente (1), e infine dall'espressione:

$$R'' = \sqrt{\frac{s}{s-1}\,\frac{m}{m-1}\left[{}^2S^2_A - p\,(1-p)\frac{m-1}{ms-1}\right]}$$

se la probabilità p viene invece desunta dalla frequenza del carattere nelle ms osservazioni (2).

Il massimo di R'' è

$$Max\,(R'') = \sqrt{\frac{s}{s-1}\,\frac{m}{m-1}\,p\,(1-p)\,\frac{m(s-1)}{ms-1}}$$

---

(1) Cfr. L. v. Bortkewitsch, *Das Gesetz der kleinen Zahlen*, Leipzig, Teubner, 1898, § 14, pagg. 29-30.

(2) Nel caso che p sia desunta dalla frequenza del carattere nelle ms osservazioni, nello scostamento quadratico medio effettivo ${}^2S_A$ bisogna introdurre il coefficiente di correzione $\sqrt{\frac{m}{m-1}}$ e, nel suo valore teorico ${}^2s_A$, il coefficiente di correzione $\sqrt{\frac{ms}{ms-1}}$, così che al valore di R' si deve sostituire il valore

$$R'' = \sqrt{\frac{s}{s-1}\left(\frac{m}{m-1}\,{}^2S_A{}^2 - \frac{ms}{ms-1}\,\frac{p\,(1-p)}{s}\right)}$$

Rapportando ad esso R'' si ottiene esattamente la (60¹.1), la quale può mettersi anche sotto la forma:

$$^2Z' = \sqrt{1 - \frac{ms - 1}{p(1-p)\, m^2 s^2} \sum_{i=1}^{m} \frac{f_i (s - f_i)}{s - 1}} \qquad (60^1.2)$$

24. Nelle applicazioni della teoria della dispersione molto spesso però gli m gruppi di osservazioni, corrispondenti ai vari termini della serie, non hanno lo stesso numero s di osservazioni. Se indichiamo con s_i il numero delle osservazioni corrispondenti al termine generico i^{mo}, con $\frac{f_i}{s_i}$ la frequenza del carattere in tali osservazioni, $p = \frac{\sum_{i=1}^{m} f_i}{\sum_{i=1}^{m} s_i}$ sarà la frequenza del carattere in tutte le $ms = \sum_{i=1}^{m} s_i$ osservazioni.

È evidentemente giustificato, in tal caso, attribuire ai quadrati degli scostamenti $\frac{f_i}{s_i} - p$, che ricorrono nel calcolo di ^2S_A, non lo stesso peso, ma un peso proporzionale al numero s_i delle osservazioni a cui la frequenza $\frac{f_i}{s_i}$ si riferisce, così che il valore di ^2S_A sarà fornito dall'espressione:

$$^2S_A = \sqrt{\frac{1}{ms} \sum_{i=1}^{m} s_i \left(\frac{f_i}{s_i} - p\right)^2}$$

Ora si dimostra (1), che, anche in tal caso, il valore di ^2s_A

(1) Vedi in proposito G. Pietra, *The theory of statistical relations with special reference to cyclical series.* «Metron», Vol. IV, N. 3-4, 1 Giugno, 1925, § 25, pag. 433-439, dove sono esposti — come l'A. debitamente ricorda — i risultati di una nostra memoria tuttora inedita. In detto paragrafo sono già contenuti quasi tutti i risultati sulle relazioni tra gli indici di dispersione e di connessione esposti in questo articolo. Vedi anche, in proposito la nota, *Su la misura della dispersione e la sua relazione con l'indice di connessione*, presentata al Seminario statistico dell'Università di Roma nella seduta del 25 aprile 1942 e pubblicata negli atti della IV e

è fornito dalla (59) e che è $Max\,(^2S_A) = \sqrt{p\,(1-p)}$, così che valgono, anche per tal caso, le (60).

La (60¹.2) non rappresenta però esattamente, in tal caso, il rapporto della componente fisica della dispersione al suo massimo, che è invece dato dalla espressione seguente (¹) :

$$^2Z' = \sqrt{1 - \frac{ms - 1}{p\,(1-p)\,m^2\,s^2} \sum_{i=1}^{m} \frac{f_i\,(s_i - f_i)}{s_i - 1}} \qquad (62)$$

mentre, in tal caso, la (60¹.2) può mettersi sotto la forma,

$$^2Z' = \sqrt{1 - \frac{ms - 1}{p\,(1-p)\,m^2\,s^2} \sum_{i=1}^{m} \frac{f_i\,(s_i - f_i)}{s_i - 1} \cdot \frac{s_i - 1}{s - 1} \frac{s - f_i}{s_i - f_i}} \qquad (60^1.3)$$

V riunione scientifica della Società Italiana di Statistica, Roma, 1942, pagg. 467-470.

Detti risultati si riferiscono però sempre a serie di frequenze relative : una generalizzazione a serie di grandezze estensive, trovasi nell'articolo *Intensità e precisione nella connessione statistica*, « Statistica », fascicolo 3-4, 1947.

Questa generalizzazione si può fare quando le grandezze estensive sono rappresentate da medie e supposto che si conoscano le distribuzioni del complesso di tutte le ms grandezze di cui le s medie risultano e, per la determinazione esatta della componente fisica della dispersione conforme alla (62), anche la distribuzione delle grandezze per ciascuna delle s medie.

(¹) In tal caso infatti, è, nei dati osservati,

$$\sum_{i=1}^{m} s_i \left(\frac{f_i}{s_i} - p\right)^2 = m\,s\,p\,(1-p) - \sum_{i=1}^{m} \frac{f_i\,(s_i - f_i)}{s_i}$$

mentre, in un numero infinitamente grande di osservazioni, sarebbe da attendersi

$$\sum_{i=1}^{m} s_i \left(\frac{f_i}{s_i} - p\right)^2 = \frac{m^2\,s^2}{ms - 1}\,p\,(1-p) - \sum_{i=1}^{m} \frac{f_i\,(s_i - f_i)}{s_i - 1}$$

La componente fisica della dispersione risulta quindi

$$R'' = \sqrt{\frac{m\,s}{ms - 1}\,p\,(1-p) - \frac{1}{m\,s} \sum \frac{f_i(s_i - f_i)}{s_i - 1}}$$

e il suo massimo, che viene raggiunto quando è sempre $f_i = 0$ oppure $f_i = s_i$, è

$$Max\,(R'') = \sqrt{\frac{m\,s}{m\,s - 1}\,p\,(1-p)}$$

la quale differisce dalla precedente per il coefficiente $\frac{s_i - 1}{s - 1} \frac{s - f_i}{s_i - f_i}$ che generalmente si può trascurare.

Nel caso particolare in cui è, per tutti i termini, $s_i = s$ la la (62) e la (60¹.3) coincidono.

25. È interessante notare come la (62) altro non sia chel'espressione dell'indice quadratico di connessione sistematica delle medie del carattere A alle modalità di un altro carattere B secondo le quali si sono distinti i termini della serie di cui si misura la dispersione, quando A sia un carattere qualitativo che assume due sole modalità o un carattere quantitativo che assume le due sole intensità 1 e 0. La (60¹.2) è l'espressione di tale indice nel caso particolare in cui tutte le distribuzioni parziali di A corrispondenti alle varie modalità di B hanno lo stesso numero di osservazioni.

L'indice quadratico di dispersione del carattere A desunto dal rapporto della componente fisica della dispersione al suo massimo, rientra quindi come un caso particolare, nell'indice quadratico di connessione sistematica. (1)

26. L'indice quadratico di dispersione, di cui nei precedenti numeri, va esente da alcuni inconvenienti che presentano le consuete misure della dispersione. Sono queste, come è noto, il coefficiente di divergenza e la componente fisica della dispersione.

Per mettere in luce tali inconvenienti, consideriamo il caso in cui si conosca esattamente la probabilità p del carattere e in cui il numero delle osservazioni sia $= s$ per tutti i termini della serie.

Il coefficiente di divergenza è allora dato dall'espressione:

$$Q = \sqrt{\frac{s^2 S_A^2}{p\,(1 - p)}}$$

---

(1) Per gli indici di connessione concreta e sistematica, vedansi la nostra memoria *Di una misura della dissomiglianza*, ecc., già citata a pag. 4. e l'altra *Nuovi contributi alla teoria delle relazioni statistiche*, in « Atti del R. Istituto Veneto di scienze, lettere ed arti », Anno acc. 1914-15, Tomo LXXIV, parte II, nonchè i corsi di lezioni citati nella nota (1) a pp. 3-4. Sulle relazioni tra indice di dispersione e indice di connessione sistematica, vedansi gli scritti citati nella nota (1) a pp. 68-69.

che, indicando con p_i la probabilità del carattere nel gruppo i, può venire posto sotto la forma

$$Q = \sqrt{1 + (s-1)\frac{\sum_{i=1}^{m}(p_i - p)^2}{p(1-p)}},$$

la quale mette in evidenza come il coefficiente di divergenza a parità della probabilità p del carattere e della espressione $\sum_{i=1}^{m}(p_i - p)^2$ (che è proporzionale alla componente fisica della dispersione), cresce col crescere del numero s delle osservazioni a cui si riferiscono i termini della serie.

Questo inconveniente, già segnalato dal Lexis, si può mettere in evidenza per altra via, richiamando che è $Max\,(^2S_A) = \sqrt{p(1-p)}$ e quindi $Max\,(Q) = \sqrt{s}$. Il massimo del coefficiente di divergenza cresce dunque col crescere del detto numero s delle osservazioni.

A misura della dispersione, il Lexis ha pertanto proposto la componente fisica della dispersione, data dall'espressione

$$R' = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(p_i - p)^2}$$

che, come si è detto, può mettersi sotto la forma

$$R' = \sqrt{\frac{s}{s-1}\left(^2S_A^2 - \frac{p(1-p)}{s}\right)}$$

Anche questo indice, però, non va scevro da inconvenienti.

Il suo valore è infatti massimo quando è massimo ^2S_A vale a dire quando è $^2S_A = \sqrt{p(1-p)}$.

Si ha allora

$$Max\,(R') = \sqrt{p(1-p)}$$

che può anche mettersi sotto la forma

$$Max\,(R') = \sqrt{\frac{1}{4} - K^2}$$

dove K esprime la differenza tra la probabilità p del carattere e $\frac{1}{2}$.

Il massimo della componente fisica varia dunque col variare della probabilità: esso presenta il suo valore più alto $= \frac{1}{2}$ per $p = \frac{1}{2}$ e valori tanto più bassi quanto più il carattere è raro o viceversa quanto più la sua frequenza si avvicina all'unità. (1)

Al fine di eliminare questo inconveniente, conviene ragguagliare la componente fisica al suo massimo, nel qual caso si perviene appunto agli indici quadratici di dispersione da noi proposti, forniti dalle espressioni (59), (60), (61).

RIASSUNTO

27. Utilizzando gli indici di dissomiglianza da noi introdotti nella metodologia statistica molti anni or sono e già posti a base della teoria della connessione, abbiamo proceduto ad un riesame della teoria degli indici di asimmetria e di curtosi che ci ha portato alla loro integrazione con altri indici ed alla loro coordinazione con gli indici di dispersione.

Abbiamo distinto l'*asimmetria* dalla *dissimmetria*, di cui la prima tiene conto solo dell'intensità e la seconda anche del verso delle differenze tra i due rami delle curve di graduazione posti a destra e a sinistra della mediana. Di entrambi gli aspetti abbiamo proposto indici assoluti e indici relativi, ottenuti, questi ultimi, ragguagliando gli indici assoluti al rispettivo massimo. Come massimo, abbiamo considerato, sia il massimo assoluto, rappresentato dalla media aritmetica, sia il massimo relativo, rappresentato dall'indice di variabilità della distribuzione. Degli

(1) Non è quindi da meravigliarsi se, adottando come indice di dispersione la componente fisica o il coefficiente di divergenza, si siano ottenuti valori particolarmente bassi per i fenomeni rari: ciò dipende certo in parte — e può dipendere unicamente — dal fatto che tali misure sono, come è spiegato nel testo, difettose. L'argomento sarebbe pertanto da riprendere, usando, per la misura della dispersione, le formule meglio rispondenti indicate nel testo e già pubblicate, del resto, negli scritti citati nella nota (1) a pp. 68-69.

indici, a cui siamo pervenuti, abbiamo illustrato le relazioni con quelli comunemente in uso.

Con procedimento parallelo, abbiamo studiato l'*anormalità* e la *disnormalità* delle distribuzioni statistiche, di cui la prima tiene conto dell'intensità delle differenze tra la curva di graduazione osservata e la rispettiva curva normale e la seconda anche del verso di tali differenze. Anche di tali aspetti delle distribuzioni abbiamo determinato sia indici assoluti che indici relativi, ottenuti, questi ultimi, ragguagliando gli indici assoluti al loro massimo relativo rappresentato dall'indice di variabilità. Per le distribuzioni di grandezze intensive, si può però ragguagliare l'indice di anormalità assoluta anche ad un massimo assoluto, che è funzione della frequenza media. Si è passati con questi indici alla teoria della *dispersione*, di cui si sono pure proposti indici appropriati, mostrando le loro relazioni con gli indici finora in uso e i vantaggi che di fronte a questi presentano. Si è anche mostrato come l'indice quadratico di dispersione proposto rientri, nel caso di dispersione iperbinomiale, come un caso particolare, negli indici quadratici di connessione.

Si è in particolare considerato quell'aspetto dell'anormalità che è chiamata di solito *curtosi*, o *eccessività*, o *iperbinomialità* e che noi abbiamo preferito chiamare *ipernormalità*.

Gli indici di anormalità e di ipernormalità, oltre che per la curva complessiva, si sono determinati anche per i suoi due rami, e si è quindi proceduto a misurare la *prevalenza dell'anormalità o ipernormalità* del ramo destro o del sinistro della curva.

L'asimmetria, la dissimmetria (o scentramento), l'anormalità, la disnormalità, la dispersione, la connessione, la ipernormalità, la prevalenza, vengono in tal modo a costituire altrettanti aspetti particolari della dissomiglianza tra le distribuzioni statistiche, ciascun aspetto essendo dotato di suoi indici particolari logicamente stabiliti e tutti mutuamente connessi, così da conferire alla materia una sistemazione razionale ed organica.

ZUSAMMENFASSUNG

Unter Benutzung der von uns vor vielen Jahren in die statistische Methodik eingeführten und bereits der Zusammenhangstheorie zugrunde gelegten Unähnlichkeitsindices wird die Theorie der Asymmetrie (Schiefe-) und Excess-Masse einer erneuten Prüfung unterzogen ; das führt dazu, sie durch andere Indices zu ergänzen und den Dispersionsindices zuzuordnen.

Die *Asymmetrie* wird von der *Dissymmetrie* (oder *Dezentrierung*) unterschieden ; erstere trägt nur der Stärke der Unterschiede zwischen den beiden rechts und links vom Zentralwert gelegenen Zweigen der Summenkurven (Graduationskurven) Rechnung, letztere auch dem Vorzeichnen derselben. Für beide Gesichtspunkte werden absolute Indices vorgeschlagen und relative Indices, welch letztere sich durch Beziehung der absoluten Indices auf ihr jeweiliges Maximum ergeben. Als Maximum wird sowohl das « absolute Maximum », dargestellt durch das arithmetische Mittel, als auch das « relative Maximum », dargestellt durch das Streuungsmass der Verteilung, betrachtet. Ferner werden die Beziehungen erläutert, die zwischen den hier erzielten Indices und den allgemein gebräuchlichen bestehen.

Durch analoges Vorgehen wird die *Anormalität* und die *Disnormalität* statistischer Verteilungen untersucht, wobei erstere dem Absolutwert der Differenzen zwischen der beobachteten Summenkurve und der Summenkurve der entsprechenden Normalkurve, letztere auch dem Vorzeichen dieser Differenzen Rechnung trägt. Auch für diese Aspekte der Verteilungen werden sowohl absolute als auch relative Indices bestimmt ; letztere erhält man, indem man die absoluten Indices durch ihr relatives Maximum, dargestellt durch das Streuungsmass, dividiert. Für die Verteilungen intensiver Grössen (Häufigkeitsverteilungen) kann man jedoch den absoluten Anormalitätsindex auch auf ein absolutes Maximum beziehen, das Funktion der mittleren Häufigkeit ist. Mit diesen Indices gelangt man zur *Dispersionstheorie*, für die ebenfalls besondere, geeignete Indices vorgeschlagen werden, deren Beziehungen zu den bisher gebräuchlichen Indices als auch ihre Vorteile den letzteren gegenüber aufgezeigt werden. Es wird auch gezeigt, dass im Falle von hyperbinomialer Dispersion der vorgeschlagene quadratische Index einem besonderen Fall von quadratischem Konnexionsindex darstellt.

Insbesondere wird dann der üblicher Weise mit *Curtosis*, *Excess* oder *Hyperbinomialität* bezeichnete Aspekt der Normalität betrachtet, für den wir die Bezeichnung *Hypernormalität* vorziehen.

Die Anormalitäts- und Hypernormalitätsindices werden ausser für die Gesamtkurve auch noch für ihre beiden Zweige bestimmt ; schliesslich wird dazu übergegangen, das *Überwiegen* (Prävalenz) *der Anormalität oder Hypernormalität* des rechten oder linken Kurvenzweiges zu messen.

Asymmetrie, Dissymmetrie (oder Dezentrierung), Anormalität, Disnormalität, Dispersion, Konnexion, Hypernormalität, Prävalenz bilden auf diese Weise ebensoviele besondere Aspekte der Unähnlichkeit zwischen statistischen Verteilungen, jeder durch seine besonderen, logisch fundierten Indices gekennzeichnet, die, alle mit einander verknüpft, so dem Stoff eine vernünftige und organische Ordnung verleihen.

RÉSUMÉ

En utilisant les indices de dissemblance, que nous avons introduits, il y a plusieurs années, dans la méthodologie statistique et que nous avons placés à la base de la théorie de la connexion, nous avons procédé à un nouvel examen de la théorie des indices d'asymétrie et de *kurtosis*, lequel nous a amenés à leur intégration au moyen d'autres indices et à leur coordination avec les indices de dispersion.

Nous avons distingué l'*asymétrie* de la *dissymétrie* (ou *décentrement*) desquelles la première tient compte seulement de l'intensité et la seconde aussi du sens des différences entre les deux branches de la courbe de graduation à droite et à gauche de la médiane. De l'un et l'autre de ces deux aspects nous avons proposé des indices absolus et des indices relatifs, obtenus, ces derniers, en rapportant les indices absolus à leur maximum respectif. Comme maximum nous avons considéré soit le maximum absolu, représenté par la moyenne arithmétique, soit le maximum relatif, représenté par l'indice de variabilité de la distribution. Des indices auxquels nous sommes parvenus nous avons illustré les rapports avec ceux qui sont ordinairement usités.

En suivant un procédé parallèle, nous avons étudié l'*anormalité* et la *disnormalité* des distributions statistiques ; la première tient compte de l'intensité des différences entre la courbe de graduation observée et la courbe normale respective, et la seconde, aussi du sens de ces différences. De ces aspects des distributions nous avons également déterminé des indices, soit absolus, soit relatifs, obtenus, ces derniers, en rapportant les indices absolus à leur maximum relatif, représenté par l'indice de variabilité. Pour les distributions de grandeurs intensives, on peut, cependant, rapporter l'indice d'anormalite absolue aussi à un maximum absolu, qui est fonction de la fréquence moyenne. Par ces indices nous avons passé à la théorie de la *dispersion*, dont nous avons proposé des indices appropriés, en montrant leurs rapports avec les indices jusqu'ici usités.

Nous avons, ensuite, considéré en particulier cet aspect de l'anormalité qui est ordinairement appelé *kurtosis*, ou *excessivité*, ou *hyperbinomialité*, et que nous avons préféré appeler *hypernormalité*.

Les indices d'anormalité et d'hypernormalité, outre que pour la courbe dans son ensemble, ont été déterminés aussi pour ses deux branches, et nous avons ensuite procédé à mesurer la *prévalence de l'anormalité ou hypernormalité* de la branche droite ou de celle gauche de la courbe.

L'asymétrie, la dissymétrie (ou décentrement), l'anormalité, la disnormalité, la dispersion, la connexion, l'hypernormalité, la prévalence, constituent ainsi tout autant d'aspects particuliers de la dissemblance entre les distributions statistiques, chacun de ces aspects étant pourvu de ses indices particuliers logiquement établis et tous mutuellement liés, de sorte à systématiser la matière de façon rationnelle et organique.

SUMMARY

Using the indices of dissimilarity which we introduced many years ago into statistical methodology and which we have already placed at the basis of the theory of connexion, we have reexamined the theory of the indices of asymmetry and kurtosis and this has led us to complete them with other indices and to coordinate them with the indices of dispersion.

We have drawn a distinction between *asymmetry* and *dissymmetry* (or *decentralisation*), the former only taking into account the intensity, while the latter also considers the sign of the differences between the two branches of the graduation curves to the right and to the left of the mean. We have suggested for both aspects absolute and relative indices, the latter obtained by dividing the absolute indices by their respective maximum. We have considered as maximum both the absolute maximum represented by the arithmetic mean, and the relative maximum represented by the index of variability of the distribution. We have illustrated the relations between the indices thus obtained and those commonly in use.

On parallel lines, we have studied the *abnormality* and the *disnormality* of statistical distributions ; the former takes into account the intensity of the differences between the graduation curve studied and the respective normal curve, while the latter also takes into consideration the sign of such differences. For these aspects of the distribution we have also determined both absolute and relative indices, the latter being obtained by dividing the absolute indices by their maximum as represented by the respective index of variability. As regards the distribution of relative frequencies, the absolute index of abnormality can be divided by an absolute maximum, which is a function of average frequency. These indices lead us to the theory of *dispersion*, for which appropriate indices are also suggested, showing their relation to those already in use.

Special consideration has also been given to that aspect of abnormality generally referred to as *kurtosis*, or *excess*, or *hyperbinomiality*, and which we prefer to call *hypernormality*.

The indices of abnormality and hypernormality have been calculated, not only for the whole curve but also for its two branches, and we have theu measured the *prevalence of the abnormality or hypernormality* of the right or left branch of the curve.

Asymmetry, dissymmetry (or decentralisation), abnormality, disnormality, dispersion, connexion, hypernormality, prevalence, are considered in such wise as special aspects of the dissimilarity between the statistical distributions, each aspect having its special indices, logically determined and reciprocally connected, so as to afford a rational and systematic ordering of the subject.

GIUSEPPE POMPILJ

Sul problema inverso della ricerca statistica

Il problema inverso della ricerca statistica, cioè il problema della determinazione *a posteriori* della probabilità d'avverarsi di certi avvenimenti futuri è suscettibile di varie formulazioni, ma, nella maggior parte delle applicazioni pratiche, si presenta sotto questa forma: « sapendo che in N osservazioni un certo evento E si è presentato a volte, si vuol determinare la probabilità che in H osservazioni future esso si presenti b volte, supponendo beninteso che nel corso delle $N + H$ osservazioni la probabilità incognita di verificarsi dell'evento rimanga costante ».

La soluzione teorica di tale problema si ottiene con estrema facilità [1] mediante una formula dovuta a P. S. Laplace [2], e impropriamente chiamata « formula di Bayes », che non starò qui a riportare, in cui però figurano in modo essenziale elementi che sono praticamente del tutto incogniti al ricercatore: mi riferisco alla distribuzione delle così dette « probabilità a priori ».

Questo grave e fondamentale inconveniente non ha però arrestato i cultori della statistica che, a partire da G. Bernoulli, in varie maniere, non sempre pur troppo felici, hanno cercato di arrivare ad una risposta, sia pure approssimata, del problema in esame.

I tentativi di soluzione più conosciuti e diffusi gravitano intorno ad un procedimento, la cui paternità risale a J. Bayes [3]

---

[1] Ved. per es.: G. Castelnuovo: *Calcolo delle probabilità*, (Zanichelli, Bologna 1925-28), Vol. I cap. IX.

[2] Ved.: P. S. Laplace, *Théorie analytique des probabilités*, ristampato nel vol. VII delle opere complete.

[3] Ved.: J. Bayes, *An essay towards solving a problem in the doctrine of chances*, Philosophical Trans. 1764.

il quale, in mancanza di ulteriori informazioni introduceva la ipotesi, che poi invece prese il nome da Laplace, della eguale possibilità di tutte le probabilità a priori. Non starò qui a riportare le numerose critiche che giustamente si rivolgono a questo modo di procedere. Sta però di fatto che non sono diffusi metodi migliori per tentar d'approssimare la soluzione del problema.

Veramente altri tipi di soluzione brillanti ed efficaci si trovano in una memoria di C. Gini (4) del 1911 che, soffocata però negli « Studi economico-giuridici della R. Università di Cagliari », non ha avuto alcuna diffusione nel pubblico degli studiosi tanto che non ho trovato affatto traccia di quei metodi nei trattati e nelle memorie che ho avuto occasione di consultare, salvo che in una nota di C. Gini e G. Livada (5) del 1943 dove però, tra i vari metodi proposti nella citata memoria di C. Gini, si fa riferimento a quello meno interessante, proposto allora solo incidentalmente dall'Autore. Del resto tale memoria è oggi praticamente introvabile ed io ne sono venuto a conoscenza solo perchè l'Autore me ne ha gentilmente data una copia dattiloscritta ! (6)

Nella memoria in questione, oltre ad un metodo, che chiamerò « ipotesi di Gini », consistente nell'assegnare alla distribuzione incognita delle probabilità a priori una forma che, pur permettendo di eseguire i calcoli con la stessa semplicità che s'incontra applicando la così detta ipotesi di Laplace, consente, attraverso il calcolo di opportuni parametri, una maggiore capacità d'adattamento ai problemi concreti in esame, oltre all'ipotesi di Gini, dicevo, vengono presentati nella sopraddetta memoria due brillantissimi metodi di soluzione chiamati dall'Autore medesimo : « metodo dei risultati diretti » e « metodo dei risultati indiretti ». Questi due metodi consistono nello scrivere in un'altra forma opportuna la formula che risolve il problema inverso, in

---

(4) Ved. : C. Gini, *Considerazioni sulle probabilità a posteriori e applicazioni al rapporto dei sessi nelle nascite umane*. Studi Economico-giuridici della R. Università di Cagliari, anno III, 1911.

(5) Ved. : C. Gini-G. Livada *Sulla probabilità inversa nel caso di grandezze intensive ed in particolare sulle sue applicazioni a collaudi per masse a mezzo di campioni*. Atti della VII riunione scientifica della Società Italiana di Statistica, 1943.

(6) Sono lieto di aggiungere, durante la correzione delle bozze, che la citata memoria del Prof. Gini è stata ristampata su Metron (1948).

modo che in essa figurino le probabilità di certi eventi di cui, attraverso adatte esperienze, si può conoscere una frequenza; dopo di che, introdotta l'ipotesi, necessaria in tutte le ricerche del genere, che le frequenze diano dei valori ben approssimati delle probabilità, si sostituiscono, nelle formule, alle probabilità teoriche le frequenze sperimentali ottenendone così un valore approssimato, che, in moltissimi casi concreti, risulta utilissimo.

Per il problema inverso della ricerca statistica nel caso di più eventi E_i ben poco si è fatto (7), perciò credo sia non privo d'interesse anche solo raggiungere l'estensione a questo caso generale, come, dietro consiglio del Prof. Gini, faccio in questa memoria, dei vari tentativi di soluzione proposti per il caso semplice di un solo evento E; tanto più che, generalizzando in tal modo il problema, meglio se ne coglie il suo contenuto astratto.

Comunemente si considera il così detto « problema inverso » come inverso di uno schema bernoulliano d'estrazione, mentre poi, quando se ne cerca la soluzione teorica, si immagina uno schema diretto che si può considerare come la generalizzazione di uno schema studiato, da P.S. Laplace; e dall'economista tedesco W. Lexis (8) e che si potrebbe senz'altro chiamare schema di Laplace. Ritengo perciò che sia ai fini pratici più opportuno contrapporre al « problema inverso » non tanto lo schema di Bernoulli quanto quello di Laplace.

Per questo motivo inizio la ricerca con una rapida indagine sulla variabile casuale delle prove ripetute a parecchie alternative secondo lo schema di Laplace (n. 1) dopo di che (n. 2) cerco di precisare la natura del problema inverso, nel caso di più eventi, ricordandone la soluzione teorica. Successivamente (nn. 3, 4) discuto i due metodi di soluzione che si ottengono generalizzando l'ipotesi detta di Laplace e quella di Gini; ed infine (nn. 5,6) ottengo la generalizzazione dei due metodi dei risultati diretti ed indiretti.

Prima di chiudere queste poche righe introduttive voglio fissare alcune convenzioni di scrittura a cui mi attengo nel seguito:

---

(7) Ved.: C. Jordan, *Inversione della formula di Bernoulli relativa al problema delle prove ripetute a più variabili*. Gior. dell'Ist. Italiano degli Attuari, 1933.

(8) Ved.: W. Lexis, *Ueber die Theorie der Stabilität Statisticher Reihen* Conrad's Jarbüchern Bd. XXXII, 1879.

— invece di scrivere « variabile casuale » siglo « v. c. »

— indico le v. c. con carattere ronde maiuscolo ; per esempio : $\mathcal{S}$

— uso il simbolo combinatorio $\binom{N}{a_1\, a_2\, \ldots\, a_k}$ (dove però $N = \sum_1^k a_i$) per indicare la quantità : $\dfrac{N!}{a_1!\, a_2! \ldots a_k!}$.

§ 1. — Posizione del problema.

1. – *Schema di* Laplace — Per evitare inutili lungaggini d'esposizione, dirò brevemente che certi k eventi E_1, $E_2, \ldots, E_k$ sono *incompatibili* e *necessari* se, in un prefissato esperimento, si presenta necessariamente uno ed uno solo di essi.

Siano dunque $E_1, E_2, \ldots, E_k$ certi k eventi incompatibili e necessari e siano inoltre $p_1, p_2, \ldots p_k$ $(\sum_1^k p_i = 1)$ le probabilità che essi hanno di presentarsi, probabilità che non saranno necessariamente fisse ma libere di variare descrivendo una *v. c.* multipla $(\mathcal{S}_1, \mathcal{S}_2, \ldots \mathcal{S}_k)$ detta *v. c. delle probabilità a priori,* le cui k componenti semplici soddisfano ciascuna alla nota limitazione $0 \leq \mathcal{S}_i \leq 1$, oltre ad essere linearmente legate tra di loro dalla relazione identica :

$$\mathcal{S}_1 + \mathcal{S}_2 + \ldots + \mathcal{S}_k = 1,$$

di modo che una qualunque di esse può essere espressa come funzione delle rimanenti $k - 1$. Possiamo perciò limitarci a considerare solo la v. c. $(k - 1)$ — pla $(\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{k-1})$, benchè assai spesso, allo copo d'ottenere semplificazioni di scrittura e di linguaggio, introdurrò, nelle formule e negli enunciati che incontreremo nel seguito, la v. c. $\mathcal{S}_k$ e la sua determinazione p_k rispettivamente in luogo della v. c. : $1 - \mathcal{S}_1 - \ldots - \mathcal{S}_{k-1}$ e della quantità : $1 - p_1 - \ldots - p_{k-1}$.

La v. c. $(k - 1)$ — pla $(\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{k-1})$ sarà nel seguito supposta continua ed avente $\omega(p_1, \ldots, p_{k-1})$ come funzione di densità. Tale v. c. si può al solito distendere su uno spazio a $k - 1$

dimensioni S_{k-1} [9], ma il suo campo di definizione è limitato al solo k — *edro*, che indicherò con T_k, delimitato dai k iperpiani: $p_1 = 0, p_2 = 0, \ldots, p_k = 0$; e questo, come è noto, significa che fuori di T_k la funzione $\omega(p_1, \ldots, p_{k-1})$ si può considerare, per definizione, costantemente nulla.

Ciò posto si immagini di eseguire sugli eventi $E_1, E_2, \ldots, E_k$ un'esperienza basata su uno schema così definito: anzitutto si ottenga, con opportuni procedimenti, una determinazione $(p_1, p_2, \ldots, p_k)$, variabile con l'esperienza medesima, della v.c. delle probabilità a priori e poi, ferma restando tale determinazione, si esegua un gruppo di N esperimenti secondo lo schema di Bernoulli; si costaterà allora che l'evento E_1 si è presentato c_1, volte, l'evento E_2 c_2 volte e così via, dove ovviamente $\sum_1^k c_i = N$.

Tale procedimento generalizza uno schema già studiato da P. S. Laplace [10] e perciò lo indicherò nel seguito con il nome di *schema di* Laplace.

Al variare dell'esperienza in esame, rimanendo però sempre eguale ad N il numero degli esperimenti che di volta in volta si fanno, la k — pla ordinata $(c_1, c_2, \ldots, c_k)$, che chiamerò *risultato* dell'esperienza medesima, descrive una v. c. $(\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k)$ che sarà chiamata *v. c. delle prove ripetute a parecchie alternative secondo lo schema di* Laplace o, più brevemente, *v. c. di* Laplace.

È necessario soffermarsi a studiare tale v. c.

Intanto le k v. c. componenti non sono indipendenti ma, oltre all'obbligo di assumere valori interi non negativi, debbono anche soddisfare alla relazione lineare identica:

$$\sum_1^k \mathcal{C}_i = N$$

di modo che potremmo limitarci, volendo, a considerare solo la v. c. $(k-1)$ — pla $(\mathcal{C}_1, \mathcal{C}_2, \ldots \mathcal{C}_{k-1})$.

Dalle cose dette segue che la v. c. di Laplace in esame è suscettibile solo di $\binom{N+k}{k}$ determinazioni, tante essendo le ma-

---

[9] Ved. per es.: G. Castelnuovo, loc. cit. (1) e G. Pompilj, *Sulla regressione*, Rend. di mat. e delle sue appl. 1946.

[10] Ved.: P. S. Laplace loc. cit. (2).

niere diverse di scomporre il numero N nella somma di k numeri interi non negativi. Inoltre la probabilità del risultato $(c_1, c_2, \ldots, c_k)$ è data ovviamente da:

$$(1) \qquad P_{c_1 \ldots c_k} = \binom{N}{c_1 \ldots c_k} M(P_1^{c_1} \ldots P_k^{c_k}) =$$

$$= \binom{N}{c_1 \ldots c_k} \int_{T_k} \ldots \int \omega(p_1 \ldots p_{k_1}) \, p_1^{c_1} \ldots p_k^{c_k} \, dp_1 \ldots dp_{k-1}$$

Un po' più laborioso sarà il calcolo dei *momenti* della v. c. in esame. Per evitare calcoli inutilmente laboriosi conviene ricercare in luogo dei « momenti » $m_{r_1 \ldots r_k}$, i così detti « *momenti fattoriali:* $m_{(r_1 \ldots r_k)}$ » definiti da:

$$(2) \qquad m_{(r_1 \ldots r_k)} = M[\mathcal{C}_1 (\mathcal{C}_1 - 1) \ldots . (\mathcal{C}_1 - r_1 + 1) \ldots .$$
$$\mathcal{C}_k (\mathcal{C}_k - 1) \ldots . (\mathcal{C}_k - r_k + 1)]$$

La convenienza a calcolare i momenti fattoriali anzichè gli usuali momenti risiede nel fatto che la funzione generatrice dei momenti fattoriali [11] $N(\beta_1 \ldots \beta_k)$, nel caso in esame, assume una forma particolarmente semplice; infatti si ha:

$$N(\beta_1, \beta_2, \ldots, \beta_k) = M[(1 + \beta_1)^{\mathcal{C}_1} (1 + \beta_2)^{\mathcal{C}_2} \ldots (1 + \beta_k)^{\mathcal{C}_k}] =$$

$$= \sum_{\Sigma r \leq N} \frac{\beta_1^{r_1} \beta_2^{r_2} \ldots \beta_k^{r_k}}{r_1! \, r_2! \ldots r_k!} \, m_{(r_1 \ldots r_k)}$$

e nel caso nostro:

$$N(\beta_1 \ldots \beta_k) = \int_{T_k} \ldots \int \omega(p_1 \ldots p_{k-1}) \left[\sum \binom{N}{c_1 \ldots c_2} p_1^{c_1} (1 + \beta_1)^{c_1} \ldots \right.$$

$$\left. \ldots p_k^{c_k} (1 + \beta_k)^{c_k} \right] dp_1 \ldots dp_{k-1} = \int_{T_k} \ldots \int \omega(p_1 \ldots p_{k-2}) (1 + p_1 \beta_1 +$$

$$+ \ldots + p_k \beta_k)^N \, dp_1 \ldots dp_{k-1} = \sum_{\Sigma r \leq N} \frac{\beta_1^{r_1} \ldots \beta_k^{r_k}}{r_1! \ldots r_k!} \, N(N-1) \ldots (N +$$

$$- r_1 - \ldots - r_k + 1) \, M(\mathcal{S}_1^{r_1} \mathcal{S}_2^{r_2} \ldots \mathcal{S}_k^{r_k}),$$

---

[11] Ved. per es.: A. C. AITKEN, *Statistical Mathematics*, Oliver & Boyd, Londra, 1942.

da cui, confrontando con la formula precedente, si ottiene:

$$(3) \quad m_{(r_1 r_2 \dots r_k)} = N\ (N-1)\ \dots\ (N - r_1 - r_2 - \dots + \\ - r_k + 1)\ M\ (\mathfrak{S}_1^{r_1}\ \mathfrak{S}_2^{r_2} \dots \mathfrak{S}_k^{r_k})\ .$$

Dalla (3) si ricavano in particolare:

$$m_{(10\dots0)} = NM\ (\mathfrak{S}_1)$$
$$m_{(110..0)} = N\ (N-1)\ M\ (\mathfrak{S}_1\ \mathfrak{S}_2)$$
$$m_{(20\dots0)} = N\ (N-1)\ M\ (\mathfrak{S}_1^2)$$
$$\dots\dots\dots\dots\dots\dots\dots\dots$$

Il passaggio dai momenti fattoriali ai momenti propriamente detti si ottiene con formule ricorrenti semplicissime a ricavarsi e praticamente note; in particolare si ha:

$$m_{10\dots0} = m_{(10\dots0)}$$
$$m_{110..0} = m_{(110..0)}$$
$$m_{20\dots0} = m_{(20\dots0)} + m_{(10\dots0)}$$
$$\dots\dots\dots\dots\dots\dots\dots\dots$$

di modo che possiamo senz'altro scrivere:

$$M\ (\mathfrak{C}_i) = N\ M\ (\mathfrak{S}_i)$$
$$M\ (\mathfrak{C}_i\ \mathfrak{C}_j) = N\ (N-1)\ M\ (\mathfrak{S}_i\ \mathfrak{S}_j)$$
$$M\ (\mathfrak{C}_i^2) = N\ (N-1)\ M\ (\mathfrak{S}_i^2) + N\ M\ (\mathfrak{S}_i)$$
$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$

Indicando ora con π_i e $\mu^2_{\mathfrak{S}_i}$ la media e la varianza della v. c. $\mathfrak{S}_i$ possiamo calcolare la varianza $\mu^2_{\mathfrak{C}_i}$ della v. c. $\mathfrak{C}_i$, ottenendo, come del resto era noto, la relazione:

$$\mu^2_{\mathfrak{C}_i} = N\ \pi_i\ (1 - \pi_i) + N\ (N-1)\ \mu^2_{\mathfrak{S}_i}$$

2. – *Problema diretto e problema inverso.* — Se al momento di eseguire le N prove avessino già conosciute le probabilità $(p_1, p_2, \dots, p_k)$ prescelte, per calcolare la probabilità del risultato $(a_1, a_2, \dots, a_k)$ $\left[\sum_i^k a_i = N\right]$ non avremmo dovuto ricorrere

ad una formula analoga alla (1), ma avremmo dovuto senz'altro scrivere [12]:

$$(4) \qquad P_{a_1 \dots a_k} = \binom{N}{a_1 \dots a_k} p_1^{a_1} \dots p_k^{a_k}$$

In queste condizioni, ferme restando le probabilità $p_1, p_2, \dots, p_k$, la k — *pla* ordinata $(a_1, a_2, \dots, a_k)$, al variare del gruppo di N prove, descrive una v. c. $(\mathfrak{A}_1, \mathfrak{A}_2, \dots, \mathfrak{A}_k)$ detta *v. c. delle prove ripetute a parecchie alternative secondo lo schema di* BERNOULLI o, più brevemente, *v. c. di* BERNOULLI. I momenti di tale v. c. sono già noti [13] e quindi non starò a riportarli, tanto più che non serviranno per il seguito.

La formula (4), o, in più generale, lo studio sistematico della *v. c. di* BERNOULLI risolve il così detto problema diretto della ricerca statistica il quale appunto si propone di studiare il comportamento dei risultati che si ottengono eseguendo delle esperienze composte da N esperimenti semplici, su certi k eventi $E_1, E_2, \dots, E_k$ necessari e incompatibili i quali hanno certe probabilità note e costanti $p_1, p_2, \dots, p_k$ di presentarsi.

A tale problema si suole contrapporre il così detto *problema inverso* che si propone invece, ferme restando le k frequenze assolute a_i, di determinare le probabilità a posteriori che le v. c. $\mathfrak{F}_i$ abbiano assunto valori degli intervalli infinitesimi p_i cioè: $p_i + dp_i$.

Tale probabilità a posteriori è ben nota e viene fornita dalla formula così detta di Bayes [14]:

$$(5) \qquad {}_{a_1, \dots a_k}\pi\,(p_1\, p_2 \dots p_{k-1})\, dp_1 \dots dp_{k-1} =$$

$$= \frac{\omega\,(p_1 \dots p_{k-1})\; p_1^{a_1} \dots p_k^{a_k}\, dp_1 \dots dp_{k-1}}{\int_{T_k} \dots \int \omega\,(p_1 \dots p_{k-1})\; p_1^{a_1} \dots p_k^{a_k}\, dp_1 \dots dp_{k-1}}$$

La funzione ${}_{a_1, \dots a_k}\pi\,(p_1\, p_2 \dots p_{k-1})$ così definitita è ovviamente la funzione di densità di un'altra v. c. che chiamerò *v. c.*

---

[12] Ved. per es.: G. CASTELNUOVO, loc. cit. (1).

[13] Ved. per es.: G. POMPILJ, *Sulla media geometrica e sopra un indice di mutabilità calcolati mediante un campione*, Mem. Soc. Italiana delle Scienze detta dei XL, serie III, vol. XXVI, 1947.

[14] Ved. per es.: G. CASTELNUOVO, loc. cit. (1).

delle probabilità a posteriori e che indicherò con: $({}_{a_1 \ldots a_k}\mathcal{E}_1, {}_{a_1 \ldots a_k}\mathcal{E}_2, \ldots, {}_{a_1 \ldots a_k}\mathcal{E}_k)$; tale scrittura è però piuttosto ingombrante e perciò il più delle volte indicherò la medesima v. c. con il simbolo più semplice: $({}_a\mathcal{E}_1, {}_a\mathcal{E}_2, \ldots, {}_a\mathcal{E}_k)$.

Anche questa volta sussiste la solita relazione:

$$ {}_a\mathcal{E}_1 + \mathcal{E}_2 + \ldots + {}_a\mathcal{E}_k = 1 $$

di modo che la precedente v. c. multipla, benchè apparentemente formata da k componenti, va però considerata come $(k-1)$ —*pla*, ed anzi, in base alle convenzioni già introdotte, come coincidente con la v. c. formata dalle prime $(k-1)$ componenti. È inoltre chiaro che il campo di definizione di tale v. c. è ancora il k —*edro* T_k già considerato.

Al fine di esporre nel modo più semplice ed evidente l'intimo significato della v. c. ora introdotta delle probabilità a posteriori, credo convenga ricorrere al linguaggio delle urne perchè, pur essendo provvisoriamente costretti a rinunziare all'ipotesi della continuità, si ottiene l'enorme vantaggio di un'esposizione chiara e suggestiva.

Si disponga dunque di un gran numero Z di urne, contenenti ciascuna W palle contrassegnate con uno dei numeri $1, 2, \ldots, k$, di modo che per ogni urna restano definite le probabilità p_i $(i = 1, 2, \ldots, k)$ d'estrarre una palla contrassegnata con il numero i. Questa prima collettività di urne fornisce quella che abbiamo chiamato « v. c. delle probabilità a priori » in quanto, una volta scelta un'urna, restano fissate le probabilità $(p_1, p_2, \ldots, p_k)$ che, al variare della scelta, descrivono la v. c. $(\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_k)$.

Supponiamo ora di eseguire su ognuna della Z urne un'esperienza formata da N prove e mettiamo da parte quelle urne che hanno fornito nell'esperienza il risultato $(a_1, a_2, \ldots, a_k)$. Con le urne così messe da parte si definisce una seconda collettività che appunto ci dà in media quella v. c. che abbiamo chiamato « delle probabilità a posteriori » nel senso che se il numero iniziale Z delle urne si fosse fatto divergere all'infinito, senza però alterare la v. c. delle probabilità a priori, una volta scelta un'urna di questa seconda collettività si sarebbe ottenuta una determinazione $(p_1, p_2, \ldots, p_k)$ della v. c. $({}_a\mathcal{E}_1, {}_a\mathcal{E}_2, \ldots, {}_a\mathcal{E}_k)$.

Chiarito così il significato della v. c. delle probabilità a posteriori conviene, prima di passare ad altro argomento, soffer-

marsi a notare la semplicissima relazione che passa tra i momenti delle due v. c. delle probabilità a posteriori e a priori:

$$M\left({}_a\mathfrak{I}_1^{r_1}\ {}_a\mathfrak{I}_2^{r_2}\ldots\ {}_a\mathfrak{I}_k^{r_k}\right)=$$

$$=\frac{\int_{T_k}\ldots\int\omega\,(p_1\ldots p_{k-1})\;p_1^{a_1+r_1}\ldots p_k^{a_k+r_k}\,dp_1\ldots dp_{k-1}}{\int_{T_k}\ldots\int\omega\,(p_1\ldots p_{k-1})\,p_1^{a_1}\ldots p_k^{a_k}\,dp_1\ldots dp_{k-1}}=$$

$$=\frac{M\left(\mathfrak{I}_1^{a_1+r_1}\,\mathfrak{I}_2^{a_2+r_2}\ldots\ldots\,\mathfrak{I}_k^{a_k+r_k}\right)}{M\left(\mathfrak{I}_1^{a_1}\,\mathfrak{I}_2^{a_2}\ldots\mathfrak{I}_k^{a_k}\right)}$$

A partire dalla v. c. delle probabilità a posteriori si può definire una seconda *v. c. di* LAPLACE che permette, ritornando al linguaggio delle urne, di studiare come si distribuiscono i risultati di gruppi di H prove eseguite sopra le urne che in un precedente gruppo di N prove hanno già dato il risultato $(a_1, a_2, \ldots, a_k)$; problema questo che fornisce l'aspetto sotto cui generalmente si presentano in pratica le applicazioni del problema inverso.

Indicherò questa seconda *v. c. di* LAPLACE con il simbolo $(a_1\ldots a_k\,\mathfrak{B}_1, a_1\ldots a_k\,\mathfrak{B}_2, \ldots, a_1\ldots a_k\,\mathfrak{B}_k)$ oppure, usando una scrittura meno precisa ma anche meno ingombrante, con: $(\mathfrak{B}_1, \mathfrak{B}_2, \ldots, \mathfrak{B}_k)$ nel seguito mi atterrò a questa seconda scrittura.

La v. c. $(\mathfrak{B}_1, \mathfrak{B}_2, \ldots, \mathfrak{B}_k)$ sopra definita, anche se non comparirà in forma esplicita, rappresenterà il motivo conduttore di tutto il seguito della memoria, perciò credo convenga darle un nome e precisamete la chiamerò *v. c. dell'inversione.*

Riepilogando le cose fin qui dette possiamo proporre in modo preciso il « problema inverso » nei termini seguenti: si sia anzitutto proceduto ad una determinazione della v. c. delle probabilità a priori, fissando cioè certe probabilità p_i $(i = 1, 2, \ldots, k)$ degli eventi E_i; si supponga inoltre che tali probabilità p_i, pur essendo ormai fissate non siano a noi note; eseguiamo ora, in tali condizioni, N prove sugli eventi necessari ed incompatibili E_i ottenendo il risultato $(a_1, a_2, \ldots a_k)$; dopo di che ci proponiamo di eseguire nelle stesse condizioni (cioè con le stesse pro-

babilità p_i) altre H prove; vagliamo conoscere la probabilità ${}_{a_1 \ldots a_k}P_{b_1 \ldots b_k}$ di ottenere in queste ulteriori H prove il risultato $(b_1, b_2, \ldots, b_k)$ $\left[\sum_1^k b_i = H\right]$.

Tale probabilità, che indicherò anche con il simbolo meno ingombrante ${}_aP_{b_1 \ldots b_k}$, non è altro che la probabilità con cui la v. c. dell'inversione $(\mathfrak{B}_1, \mathfrak{B}_2, \ldots, \mathfrak{B}_k)$ assume la determinazione $(b_1, b_2, \ldots, b_k)$ e quindi, trattandosi in definitiva di una probabilità relativa ad una *v. c. di* LAPLACE, è fornita, salvo gli opportuni adattamenti, dalla (1), dove però al posto della $\omega(p_1, p_2, \ldots, p_{k-1})$ va messa la ${}_{a_1 \ldots a_k}\pi(p_1, p_2, \ldots, p_{k-1})$ fornita dalla (5); si ha pertanto:

$$(6) \qquad {}_aP_{b_1 \ldots b_k} =$$

$$= \binom{H}{b_1 \ldots b_k} \frac{\int_{T_k} \ldots \int \omega(p_1 \ldots p_{k-1})\, p_1^{a_1+b_1} p_2^{a_2+b_2} \ldots p_k^{a_k+b_k} dp_1 \ldots dp_{k-1}}{\int_{T_k} \ldots \int \omega(p_1 \ldots p_{k-1})\, p_1^{a_1} p_2^{a_2} \ldots p_k^{a_k} dp_1 \ldots dp_{k-1}}$$

Naturalmente tale formula non è utilizzabile fino a che non si conosca la funzione di densità $\omega(p_1, p_2, \ldots, p_{k-1})$, il che in pratica accade assai raramente; ed allora, se anche in tali casi si vuol cercare di dir qualche cosa, bisogna sopperire a tale lacuna delle nostre conoscenze introducendo delle ipotesi sulla natura della funzione incognita $\omega(p_1 \ldots p_{k-1})$, oppure cercando altre maniere di scrivere la (6) in modo che in essa figurino probabilità di eventi, di cui si sappia calcolare la frequenza in un certo numero di prove, in modo da poter ottenere un'espressione approssimata della (6) sostituendo alle probabilità teoriche le frequenze sperimentali; il che naturalmente non si può fare senza prima introdurre l'ipotesi che, nel particolare problema in esame, sia lecita tale sostituzione. Questi due metodi di affrontare il problema sono stati già sviluppati nel caso di $k = 2$ e, come ho avuto occasione di dire nelle precedenti pagine introduttive, il primo ha dato in particolare luogo all'ipotesi detta di Laplace [15] e a quella

---

[15] Allo scopo di evitare confusioni mi uniformerò all'uso corrente e seguiterò a parlare di ipotesi di LAPLACE, benchè la paternità di tale procedimento si debba far risalire a J. BAYES, loc. cit. (3).

di Gini ([16]), mentre il secondo ha dato luogo ai due metodi dei risultati diretti e indiretti proposti ambedue da C. Gini ([17]).

Nei due paragrafi successivi di questa memoria estenderò questi vari metodi al caso di k qualunque, trattando prima del metodo delle ipotesi (§ II) e poi del metodo dei risultati (§ III).

§ II. — Metodo delle ipotesi.

3. – *Ipotesi di* Laplace. — Nel caso di due soli eventi incompatibili e necessari, ha avuto molto successo, senza dubbio di gr n lunga superiore alla sua vera portata, un procedimento che si può far risalire, come ho già detto, a J. Bayes ma che poi è stato studiato anche da Laplace ([18]) ; tale procedimento consiste nell'introdurre, in mancanza di ulteriori informazioni sulla funzione di densità delle probabilità a priori, l'ipotesi che tale funzione sia *costante*. Si è tentato, con argomentazioni più o meno speciose, che non starò qui nè a riportare nè a confutare, di fornire una giustificazione di tale procedimento in quale forse oggi non ha altri motivi a suo favore se non la . . . nobiltà dei natali e la vetustà del lignaggio !

In ossequio a tali tradizioni farò vedere in questo numero come tale ipotesi si estende al caso di k qualunque ; fermandomi però alla fine a sviluppare un motivo che chiaramente milita contro l'impiego di tale ipotesi.

L'ipotesi di Laplace può nel caso nostro enunciarsi così : « la funzione $\omega(p_1 p_2 \dots p_{k-1})$ è costante e precisamente eguale a $(k-1)!$ » (essendo $1/(k-1)!$, come è noto, il volume del k — edro T_k).

In tali condizioni la (5) diventa :

$$(7) \qquad a_1 \dots a_k \pi(p_1 p_2 \dots p_{k-1})\, dp_1\, dp_2\, dp_{k-1} =$$

$$= \frac{p_1^{a_1} p_2^{a_2} \dots p_k^{a_k}}{\int_{T_k} \dots \int p_1^{a_1} p_2^{a_2} \dots p_k^{a_k}\, dp_1\, dp_2 \dots dp_{k-1}}\, dp_1\, dp_2 \dots dp_{k-1}$$

([16]) Ved. : C. Gini, loc. cit. (4).
([17]) Ved. : C. Gini, loc. cit. (4).
([18]) Ved. ; P. S. Laplace, loc. cit. (2).

Ricordando il noto integrale euleriano di prima specie generalizzato [19]:

$$(8) \qquad B(a_1+1,\ a_2+1,\ \ldots,\ a_k+1) =$$

$$= \int_{T_k} \cdots \int p_1^{a_1}\, p_2^{a_2} \cdots p_{k-1}^{a_{k-1}} (1 - p_1 - \ldots - p_{k-1})^{a_k}\, dp_1 \ldots dp_{k-1} =$$

$$= \frac{\Gamma(a_1+1) \ldots \Gamma(a_k+1)}{\Gamma(N+k)}$$

la (7) diventa:

$$(9) \qquad a_1 \ldots a_k \pi(p_1 \ldots p_{k-1})\, dp_1\, dp_2 \ldots dp_{k-2} =$$

$$- \frac{(N+k-1)!}{a_1! \ldots a_k!}\, p_1^{a_1}\, p_2^{a_2} \ldots p_k^{a_k}\, dp_1\, dp_2 \ldots d\, p_{k-1}$$

Nell'ipotesi di Laplace i momenti della v. c. delle probabilità a priori assumono una forma particolarmente semplice:

$$(10) \quad M(\mathfrak{Z}_1^{r_1} \mathfrak{Z}_2^{r_2} \ldots \mathfrak{Z}_k^{r_k}) = (k-1)!\, B(r_1+1, r_2+1, \ldots, r_k+1) =$$

$$= \frac{(k-1)!\, r_1!\, r_2! \ldots r_k!}{(r_1 + r_2 + \ldots + r_k + k - 1)!}$$

e di conseguenza una forma semplicissima assumono i momenti fattoriali della v. c. di Laplace:

$$(11) \qquad m_{(r_1 \ldots r_k)} = N(N-1) \ldots (N - r_1 - \ldots - r_k + 1)$$

$$\frac{(k-1)!\, r_1! \ldots r_k!}{(r_1 + r_2 + \ldots + r_k + k - 1)!}$$

Non mi soffermerò a discutere tali risultati e passo senz'altro a calcolare i momenti della v. c. delle probabilità a posteriori. In base alle formule precedenti si ha:

$$(12) \qquad M({}_a\mathfrak{Z}_1^{r_1} \ldots {}_a\mathfrak{Z}_k^{r_k}) = \frac{B(a_1 + r_1 + 1, \ldots, a_k + r_k + 1)}{B(a_1+1, \ldots, a_k+1)} =$$

$$= \frac{(N+k-1)!\, (a_1+r_1)! \ldots (a_k+r_k)!}{a_1! \ldots a_k!\, (N+R+k-1)!}$$

dove vi è posto: $R = \sum_1^k r_i$

---

[19] Ved. per es.: Wittaker-Watson, *A Course of Modern Analysis* University Press, Cambridge, 1927.

In particolare dalla (12) si ottengono le formule:

$$(13)\qquad \begin{cases} M\,({}_a\mathfrak{S}_i) = \dfrac{a_i + 1}{N + k} \\ M\,({}_a\mathfrak{S}_i\ {}_a\mathfrak{S}_j) = \dfrac{(a_i + 1)(a_j + 1)}{(N + k)\,(N + k + 1)} \\ M\,({}_a\mathfrak{S}_i^2) = \dfrac{(a_i + 1)\,(a_i + 2)}{(N + k)\,(N + k + 1)} \end{cases}$$

mediante cui si possono calcolare la varianza μ_i^2 della v. c. ${}_a\mathfrak{S}_i$ ed il coefficiente di correlazione r_{ij} della v. c. doppia $({}_a\mathfrak{S}_i\,,\,{}_a\mathfrak{S}_j)$; conviene però, al fine d'ottenere formule di facile interpretazione, indicare con π_i il valor medio della v. c. ${}_a\mathfrak{S}_i$, cioè:

$$(14)\qquad M\,({}_a\mathfrak{S}_i) = \frac{a_i + 1}{N + k} = \pi_i$$

dopo di che con facili calcoli si trova:

$$(15)\qquad \begin{cases} \mu^2{}_i = \dfrac{\pi_i\,(1 - \pi_i)}{N + k + 1} \\ r_{ij} = -\sqrt{\dfrac{\pi_i\,\pi_j}{(1 - \pi_i)\,(1 - \pi_j)}} \end{cases}$$

Sviluppando le considerazioni sopra esposte si potrebbero trovare quanti valori segnaletici si vogliono per la v. c. delle probabilità a posteriori; ma per ora mi basta aver calcolato i due precedenti parametri il cui significato è evidente.

La relazione (12) permette di risolvere, nell'ipotesi in esame, il problema enunciato al n. 2, di calcolare cioè la probabilità che dopo aver ottenuto nelle prime N prove il risultato $(a_1, a_2, \ldots, a_k)$ nelle successive H prove si ottenga il risultato $(b_1, b_2, \ldots, b_k)$.

In base alla formula (6) tale probabilità è data da

$${}_aP_{b_1 \ldots b_k} = \binom{H}{b_1\ b_2 \ldots b_k} M\,({}_a\mathfrak{S}^{b_1}\ {}_a\mathfrak{S}_2^{b_2} \ldots {}_a\mathfrak{S}_k^{b_k})$$

da cui, mediante la (12), si ottiene

$$(16) \qquad {}_aP_{b_1 \ldots b_k} = \frac{H!}{b_1! \ldots b_k!} \cdot \frac{(N+k-1)!}{a_1! \ldots a_k!} \cdot$$

$$\cdot \frac{(a_1+b_1)! \ldots (a_k+b_k)!}{(N+H+k-1)!}$$

In particolare la probabilità che, dopo i noti risultati ottenuti nelle N prove, in un esperimento successivo si ottenga lo evento E_i è data da:

$${}_aP_i = \pi_i = \frac{a_i+1}{N+k}$$

Vale poi la pena di osservare come l'ipotesi di Laplace non comporta affatto che sia costante la funzione di densità delle singole v. c. $\mathcal{S}_i$; in realtà tutte queste v. c. hanno la stessa funzione di densità che è espressa da:

$$(17) \qquad \omega(p_i) = (k-1)(1-p_i)^{k-2}$$

La formula (17) va riguardata come particolarizzazione di un risultato generale che val la pena di stabilire perchè ci permetterà di prospettare ancora un'altro motivo per sconsigliare l'uso dell'ipotesi di Laplace.

Supponiamo di considerare come unico evento E il fatto che si presenti o l'evento E_1, o l'evento $E_2, \ldots$, o l'evento E_{r+1} di modo che, dopo questo raggruppamento, gli eventi da considerare sono solo $k-r\ (>1)$ e precisamente: $E, E_{r+2}, \ldots, E_k$, i quali si presenteranno con certe probabilità $p, p_{r+2}, \ldots, p_k$ che descrivono la v. c. $(k-r)-pla\ (\mathcal{S}, \mathcal{S}_{r+2}, \ldots, \mathcal{S}_k)$ o meglio la la v. c. $(k-r-1)-pla\ (\mathcal{S}, \mathcal{S}_{r+2}, \ldots, \mathcal{S}_{k-1})$ dove, per il teorema delle probabilità totali, si ha: $\mathcal{S} = \mathcal{S}_1 + \ldots + \mathcal{S}_{r+1}$.

Se la v. c. $(\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{k-1})$ ubbidisce all'ipotesi di Laplace, indicata con $\omega(p, p_{r+2}, \ldots, p_{k-1})$ la funzione di densità della v. c. $(\mathcal{S}, \mathcal{S}_{r+2}, \ldots, \mathcal{S}_{k-1})$, si ottiene, con facili calcoli, l'identità:

$$\omega(p, p_{r+2}, \ldots, p_{k-1}) = \frac{(k-1)!}{r!} p^r$$

In particolare, se immaginiamo di raggruppare nell'evento E tutti gli eventi $E_1, E_2, \ldots$ ad eccezioni dell'evento E_i si ha:

$$\omega(p) = (k-1)\, p^{k-2}$$

e finalmente, ricordando l'identità $p + p_i = 1$, si ritrova che la funzione di densità della v. c. $\mathcal{E}_i$ è proprio data dalla (17).

Dalle cose dette segue che la nuova v. c. $(\mathcal{E}, \mathcal{E}_{r+2}, \ldots, \mathcal{E}_{k-1})$ delle probabilità a priori, relativa al raggruppamento in un'unico evento E degli $r+1$ eventi $E_1, E_2, \ldots, E_{r+1}$, non ubbidisce più all'ipotesi di Laplace. Esprimerò brevemente questo fatto dicendo che l'ipotesi di Laplace non soddisfa al *principio della conservazione formale*; il che è invero un fatto assai grave perchè, salvo casi artificialmente preparati, ciascuno degli eventi E_i può essere in pratica considerato come raggruppamento di altri eventi $E_{i_1}, E_{i_2}, \ldots$ ed è evidente che un vero procedimento generale (e così viene considerato dai suoi fautori il metodo dell'ipotesi di Laplace) deve portare a risultati concordi comunque si raggruppino gli eventi, soddisfacendo in tal modo al principio della conservazione formale!

4. – *Ipotesi di* GINI. — L'ipotesi di Laplace presenta, come si è potuto costatare il vantaggio di fornire, dal punto di vista formale, una semplicissima soluzione del problema inverso; nasce quindi spontaneo il desiderio di formulare un'ipotesi che pur conservando intatta la semplicità dei procedimenti analitici, abbia tuttavia una qualche capacità d'adattamento alla realtà del singolo problema in esame, capacità che manca del tutto all'ipotesi di Laplace.

Alle due esigenze sopra esposte soddisfa invece senza dubbio l'ipotesi introdotta da C. Gini ([20]) per il caso $k = 2$, che chiamerò «ipotesi di Gini»; senza contare poi che essa soddisfa al principio della conservazione formale.

Nel caso generale in esame, l'ipotesi di Gini si può enunciare così: *la funzione di densità della v. c. delle probabilità a priori è data da*:

$$\omega(p_1 \ldots p_{k-1}) = \frac{\Gamma(R+k)}{\Gamma(r_1+1)\ldots\Gamma(r_k+1)}\, p_1^{r_1} p_2^{r_2} \ldots p_k^{r_k} \tag{18}$$

([20]) Ved.: C. GINI, loc. cit. (4).

dove le r_i sono numeri reali non negativi ed R è uguale alla loro somma.

In queste condizioni la funzione di densità della v. c. delle probabilità a posteriori è data da:

$$(19) \qquad {}_{a_1 \ldots a_k}\pi\,(p_1 \ldots p_{k-1}) = \frac{\Gamma\,(N + R + k)}{\Gamma\,(a_1 + r_1 + 1) \ldots \Gamma\,(a_k + r_k + 1)}$$

$$p_1^{a_1+r_1}\, p_2^{a_2+r_2} \ldots p_k^{a_k+r_k}$$

dove si è usata la funzione Γ anzichè i fattoriali perchè non è a priori garantito che le r_i siano interi, anzi in generale non lo saranno; ma a prescindere da questo particolare più che trascurabile si può costatare e che la (19) coincide con la funzione che si ottiene nell'ipotesi di Laplace quando al posto di N si metta $N + R$ e al posto di $a_1, a_2, \ldots, a_k$ si sostituisca $a_1 + r_1, a_2 + r_2, \ldots, a_k + r_k$.

Questo permette di scrivere subito le formule analoghe a quelle stabilite nel numero precedente; in particolare rimangono invariate le (15) dove però questa volta si ponga:

$$\pi_i = M\,({}_a\mathcal{S}_i) = \frac{a_i + r_i + 1}{N + R + k}$$

e la formula fondamentale (16) diventa:

$$(20) \qquad {}_aP_{b_1 \ldots b_k} = \frac{H!}{b_1! \ldots b_k!} \cdot \frac{\Gamma\,(N + R + k)}{\Gamma\,(a_1 + r_1 + 1) \ldots \Gamma\,(a_k + r_k + 1)} \cdot$$

$$\cdot \frac{\Gamma\,(a_1 + r_1 + b_1 + 1) \ldots \Gamma\,(a_k + r_k + b_k + 1)}{\Gamma\,(N + R + H + k)}$$

La superiorità dell'ipotesi di Gini su quella di Laplace sta, come già si è detto, nel fatto che pur conservando, in sede d'applicazioni pratiche, la stessa semplicità di metodi e di risultati, tuttavia permette di ottenere valori più aderenti alla realtà in quanto si possono sfruttare tutti gli altri elementi di cui si è eventualmente a conoscenza, facendoli intervenire per calcolare nel modo più opportuno i k parametri r_i.

Esporrò ora un metodo per il calcolo di questi parametri che estende un altro già proposto da C. Gini e G. Livada [21] per il caso di due soli eventi necessari e incompatibili.

---

[21] Ved.: C. Gini- G. Livada, loc. cit. (5).

Si supponga che del fenomeno in esame siano state già fatte certe h rilevazioni, secondo lo schema di Laplace, eseguendo successivamente gruppi di $N_1, N_2, \ldots, N_h$ prove; di modo che si sappia che l'evento E_i nell'j-mo gruppo di prove si sia presentato $c_i^{(j)}$ volte $(i = 1, 2, \ldots, k\ ;\ j = 1, 2, \ldots, h)$.

Mediante i campioni così ottenuti possiamo calcolare la frequenza f_i, che supporremo maggiori di zero, dei singoli E_i:

$$f_i = \frac{\sum_1^h{}_j\, c_i^{(j)}}{\sum_1^h{}_j\, N_j} \tag{21}$$

valore quest'ultimo che dovrebbe essere circa eguale al valor medio della v. c. $\mathfrak{F}_i$

$$M(\mathfrak{F}_i) = \frac{r_i + 1}{R + k}$$

e quindi eguagliando le due formule precedenti possiamo scrivere

$$f_i = \frac{r_i + 1}{R + k} \tag{22}$$

Si scrivono così, tenuto conto dei k eventi E_i, k relazioni una qualunque delle quali però risulta conseguenza delle rimanenti, se si tien conto della relazione $R = \sum r_i$. Occorre quindi stabilire ancora almeno un'ultima relazione, e a questo scopo ci saranno utili i secondi momenti fattoriali già considerati al n. 1 delle v. c. $\mathfrak{C}_i$, componenti della *v. c. di* Laplace.

Si considerino infatti le h quantità

$$\frac{c_i^{(1)}\,(c_i^{(1)} - 1)}{N_1\,(N_1 - 1)},\ \frac{c_i^{(2)}\,(c_i^{(2)} - 1)}{N_2\,(N_2 - 1)},\ \ldots\ldots,\ \frac{c_i^{(h)}\,(c_i^{(h)} - 1)}{N_h\,(N_h - 1)}$$

ciascuna delle quali al variare del gruppo delle h esperienze varia in una v. c. il cui valor medio, per la formula (3), è proprio $M(\mathfrak{F}_i^2)$; possiamo allora dire che la media aritmetica ponderata di questi valori fornisce un valore approssimato del secondo momento della v. c. $\mathfrak{F}_i$; cioè:

$$\frac{\sum_1^h{}_j\, c_i^{(j)}\,(c_i^{(j)} - 1)}{\sum_1^h{}_j\, N_j\,(N_j - 1)} \simeq M(\mathfrak{F}_i^2) = \frac{(r_i + 1)\,(r_i + 2)}{(R + k)\,(R + k + 1)} \tag{23}$$

Sostituendo al segno $\simeq$ (leggi « circa eguale ») il segno di eguaglianza si ottengono, al variare dell'indice i, altre k relazioni tra i parametri r_i. Si potrebbero allora calcolare, con un qualche metodo d'adattamento, per esempio con il metodo dei minimi quadrati, le k quantità r_i cercando di soddisfare nel miglior modo possibile alle $2k - 1$ relazioni trovate. Ma d'altra parte, lavorando per approssimazione, non vale la pena di ricorrere a procedimenti complicati anche se apparentemente più esatti e ci si può limitare a sfruttare le ultime k relazioni trovate riunendole, per somma, in un'unica equazione nel modo seguente:

$$\frac{\sum_1^k{}_i \sum_1^k{}_j c_i^{(j)} (c_i - 1)}{\sum_1^k{}_j N_j (N_j - 2)} = \frac{\sum_1^k{}_i (r_i + 1)(r_i + 2)}{(R + k)(R + k + 1)}$$

Indicando il valore sperimentale del primo membro con ${}_c m_2$ la precedente relazione diventa:

$$(24) \qquad {}_c m_2 = \frac{\sum_1^k{}_i r_i^2 + 3R + 2k}{(R + k)(R + k + 1)} .$$

Riepilogando, per calcolare i k parametri r_i ci si può servire del seguente sistema d'equazioni:

$$(25) \qquad \begin{cases} f_1 = \dfrac{r_1 + 1}{R + k} \\ f_2 = \dfrac{r_2 + 1}{R + k} \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ f_k = \dfrac{r_k + 1}{R + k} \\ {}_c m_2 = \dfrac{\sum_1^k{}_i r_i^2 + 3R + 2k}{(R + k)(R + k + 1)} \end{cases}$$

dove, ripeto una qualunque delle prime k è conseguenza delle rimanenti $k - 1$ se si tiene conto che $R = \sum_{1}^{k} r_i$.

Risolvendo il sistema (25) si ottengono i seguenti valori dei parametri r_i:

$$r_i = f_i \frac{1 - {}_c m_2}{{}_c m_2 - \sum_{1}^{k} f_i^2} - 1. \tag{26}$$

Se i valori r_i così trovati sono positivi possiamo allora servirci di essi ed applicare la (20).

§ III. — Metodo dei risultati.

5. – *Metodo dei risultati diretti.* — Il metodo così detto « dei risultati » proposto da C. Gini [22] per risolvere il problema inverso della ricerca statistica nel caso di due soli eventi necessari e incompatibili, si basa essenzialmente su un'ingegnosa maniera di scrivere la formula generale (6) che risolve tale problema, in modo che in essa figurino delle probabilità di cui si riesce a determinare sperimentalmente la relativa frequenza; dopo di che, sostituendo alle probabilità le relative frequenze si possono ottenere dei valori approssimati della formula che risolve il problema.

Il metodo dei risultati, così come è stato concepito da C. Gini prevede due schemi di soluzione che furono chiamati: metodo dei risultati diretti e metodo dei risultati indiretti.

In questo numero estendo al caso generale il metodo dei risultati diretti riservandomi di estendere l'altro metodo nel numero successivo. A tale scopo converrà calcolare alcune probabilità.

Ricordo anzitutto che in base alle convenzioni di linguaggio introdotte, per dire che in certe $N + H$ prove l'evento E_1 si è presentato $a_1 + i_1$ volte, l'evento E_2 si è presentato $a_2 + i_2$ volte, ... , le evento E_k si è presentato $a_k + i_k$ volte, dove

---

[22] Ved.: C. Gini, loc. cit. (4).

$\sum_1^k{}_j\ a_j = N$ e $\sum_1^k{}_j\ i_j = H$, dirò brevemente che in dette prove si è presentato il *risultato* $(a_1 + i_1\ ,\ a_2 + i_2\ ,\ \ldots\ a_k + i_k)$.

La probabilità $P_{a_1 + i_1 \ldots a_k + i_k}$ di tale risultato è stata già calcolata al n. 1 a proposito dello schema di LAPLACE e pertanto, in base alla (1), è data da:

$$P_{a_1 + i_1 \ldots a_k + i_k} = \binom{N + H}{a_1 + i_1 \ldots a_k + i_k} \int \ldots \int_{T_k} \omega\,(p_1 \ldots p_{k-1})\, p_1^{a_1 + i_1} \ldots p_k^{a_k + i_k}\, dp_1 \ldots dp_{k-1}\,. \tag{27}$$

Il simbolo con cui è indicata tale probabilità è però un po' ingombrante e perciò il più delle volte userò quest'altra scrittura: P_{a+i}.

Occorre ora calcolare la probabilità $p(a_1, a_2, \ldots, a_k \,/\, i_1, i_2, \ldots, i_k)$ che, essendosi verificato il risultato $(a_1 + i_1\ ,\ a_2 + i_2\ ,\ \ldots\ ,\ a_k + i_k)$, nelle prime N prove si sia avuto il risultato $(a_1, a_2, \ldots, a_k)$; anche tale probabilità è indicata da un simbolo troppo ingombrante, perciò normalmente la indicherò in quest'altra maniera più concisa: $p\,(a \,/\, i)$. Per il calcolo di tale probabilità ci si può ricondurre ad un problema d'estrazione in blocco da un urna. Sia infatti U un'urna composta da $a_1 + i_1$ palle contrassegnate, per esempio, con la cifra 1 che chiamerò palle «1», da $a_2 + i_2$ palle «2», ..., da $a_k + i_k$ palle «k»; è ovvio che la probabilità $p\,(a/i)$ sopra richiesta coincide con quella d'estrarre da tale urna un blocco di N palle di cui a_1 siano «1», a_2 siano «2», ...; a_k siano «k»; si può quindi scrivere:

$$p\,(a/i) = \frac{\binom{N}{a_1 \,.\, a_k}\binom{H}{i_1 \,.\, i_k}}{\binom{N + H}{a_1 + i_1 \ldots a_k + i_k}} \tag{28}$$

La (28) è una formula nota, ma comunque si giustifica subito osservando che il denominatore ci dà il numero di tutte le possibili permutazioni che si possono fare con le $N + H$ palle dell'urna considerando però come identiche tutte le palle contrassegnate con la stessa cifra, mentre il numeratore ci dà il numero

di quelle permutazioni che in una certa posizione, per esempio nei primi N posti, posseggono a_1 palle « 1 », a_2 palle « 2 », . . . , a_k palle « k ».

Mi propongo ora di dimostrare che la probabilità ${}_aPb_1 \ldots b_k$ già fornita dalla (6) può essere posta sotto quest'altra forma:

$$(29) \qquad {}_aP_{b_1 \ldots b_k} = \frac{P_{a+b}\, p\,(a/b)}{\sum\limits_{\Sigma i = H} P_{a+i}\, p\,(a/i)}$$

dove con il simbolo $\sum\limits_{\Sigma i = H}$ si intende il sommatorio rispetto a tutte le k — ple di numeri interi non negativi $(i_1, i_2, \ldots, i_k)$ tali che: $\sum\limits_1^k i_j\, i_j = H$.

Per dimostrare la (29) basterà sostituire a secondo membro al posto delle P_{a+i} e delle $p\,(a/i)$ le loro espressioni fornite dalle (27) e (28), e verificare che così si ottiene proprio il secondo membro della (6). Si ha infatti:

$$\frac{P_{a+b}\, p\,(a/b)}{\sum\limits_{\Sigma i = H} P_{a+i}\, p(a/i)} = \frac{\binom{N}{a_1 \ldots a_k}\ \binom{H}{b_1 \ldots b_k}}{\binom{N}{a_1 \ldots a_k}} \cdot$$

$$\cdot \frac{\int \ldots \int_{T_k} \omega\,(p_1 \ldots p_{k-1})\, p_1^{a_1+b_1} \ldots p_k^{a_k+b_k}\, dp_1 \ldots dp_{k-2}}{\int \ldots \int_{T_k} \omega\,(p_1 \ldots p_{k-1}) \left[\sum\limits_{\Sigma i = H} \binom{H}{i_1 \ldots i_k} p_1^{i_1} \ldots p_k^{i_k} \right] p_1^{a_1} \ldots p_k^{a_k}\, dp_1 \ldots dp_{k-1}}$$

ma:

$$\sum_{\Sigma i = H} \binom{H}{i_1 \ldots i_k} p_1^{i_1} \ldots p_k^{i_k} = (p_1 + p_2 + \ldots + p_k)^H = 1$$

e quindi:

$$\frac{P_{a+b}\, p\,(a/b)}{\sum\limits_{\Sigma i = H} P_{a+i}\, p\,(a/i)} = \binom{H}{b_1 \ldots b_k}.$$

$$\cdot \frac{\int \ldots \int_{T_k} \omega\,(p_1 \ldots p_{k-1})\; p_1^{a_1+b_1} \ldots p_k^{a_k+b_k}\, dp_1 \ldots dp_{k-1}}{\int \ldots \int_{T_k} \omega\,(p_1 \ldots p_{k-1})\; p_1^{a_1} \ldots p_k^{a_k}\, dp_1 \ldots dp_{k-1}}$$

c. v. d.

A secondo membro della (29) figurano le probabilità P_{a+i} e $p\,(a/i)$ di cui in realtà solo le seconde sono senz'altro note attraverso le (28), mentre le probabilità P_{a+i} non possono essere considerate note se non si conosce la funzione $\omega\,(p_1 \ldots p_{k-1})$ di modo che, come del resto era da attendersi, sia sotto la forma (6) che (29) la difficoltà risiede sempre nella non conoscenza della funzione di densità della v. c. delle probabilità a priori.

Il vantaggio della (29) sulla (6) risiede però nel fatto che in alcuni casi permette di ottenere un valore ben approssimato della probabilità ${}_aP_{b_1 \ldots b_k}$ cui si arriva sostituendo alle probabilità incognite P_{a+i} le frequenze relative f_{a+i} degli eventi cui dette probabilità si riferiscono.

Se infatti sulla *v. c. di* LAPLACE si sono già eseguite Z esperienze consistenti ognuna in $N + H$ esperimenti semplici, si possono calcolare le frequenze relative f_{a+i} dei risultati $(a_1 + i_1\,,\ a_2 + i_2\,,\ \ldots,\ a_k + i_k)$ e poi, *ammesso che le frequenze così ottenute*, quando Z è sufficientemente grande, *forniscono delle buone approssimazioni delle probabilità incognite*, si può porre

$$(30)\quad {}_aP_{b_1 \ldots b_k} \simeq \frac{\binom{H}{b_1 \ldots b_k}}{\binom{N+H}{a_1+b_1 \ldots a_k+b_k}} \cdot \frac{f_{a+b}}{\sum_{\Sigma i = H} \frac{\binom{H}{i_1 \ldots i_k}}{\binom{N+H}{a_1+i_1 \ldots a_k+i_k}}\, f_{a+i}}$$

Questa è la formula che estende al caso generale il metodo di C. Gini dei risultati diretti.

Al variare del gruppo dell'Z esperienze, il secondo membro della (30) varia e descrive una v. c. il cui studio sistematico sarebbe di grande interesse in tali generi di questioni; ma su questo argomento tornerò in altra occasione.

6. – *Metodo dei risultati indiretti.* — In base alla formula (27), la probabilità del risultato $(a_1 + t_1, a_2 + t_2, \ldots, a_k + t_k)$ è data da:

$$P_{a+t} = \binom{N+T}{a_1 + t_1 \ldots a_k + t_k} \int \ldots \int_{Tk} \omega(p_1 \ldots p_{k-1})\, p_1^{a_1+t_1} \ldots p^{a_k+t_k}\, dp_1 \ldots dp_{k-1} \tag{31}$$

dove $N = \sum_1^k a_i$ e $T = \sum_1^k t_i$.

Supposto poi che sussistano le k diseguaglianze: $t_j \geqslant i_j$ $(j = 1, 2, \ldots, k)$ si ha, dalla (28), che la probabilità $p\,(a + i/t - i)$ che, essendosi verificato il risultato $(a_1 + t_1, a_2 + t_2, \ldots, a_k + t_k)$ nelle prime $N + H$ prove $(H = \sum_1^k i_j)$ delle $N + T$ si sia avuto il risultato $(a_1 + i_1, a_2 + i_2, \ldots, a_k + i_k)$ è fornita da:

$$p\,(a + i/t - i) = \frac{\binom{N+H}{a_1 + i_1 \ldots a_k + i_k} \binom{T-H}{t_1 - i_1 \ldots t_k - i_k}}{\binom{N+T}{a_1 + t_1 \ldots a_k + r_k}} \tag{32}$$

Ciò posto si controlla immediatamente che sussiste la relazione:

$$P_{a+i} = \sum_{\Sigma t = T}^{t \geqslant i} P_{a+t}\, p\,(a + i/t - i) \tag{33}$$

dove con il simbolo $\sum_{\Sigma t = T}^{t \geqslant i}$ si vuol significare che la somma va estesa a tutte le k-ple di valori interi non negativi $(t_1, t_2, \ldots, t_k)$ tali che $\sum_1^k t_i = T$ ed inoltre: $t_j \geqslant i_j$.

Per dimostrare la (33) basta sostituire a secondo membro P_{a+t} e $p\,(a + i/t - i)$ con le espressioni fornite dalla (31) e dalla (32) e controllare che così si ottiene il secondo membro della (27).

Riunendo la (29) e la (33) si ottiene ancora un'altra espressione della probabilità ${}_aP_{b_1 \ldots b_k}$ e precisamente:

$$ {}_aP_{b_1 \ldots b_k} = \frac{\sum_{\Sigma t = b}^{t \geqslant b} P_{a+t}\ p\,(a + b/t - b)\ p\,(a/b)}{\sum_{\Sigma i = H} \sum_{\Sigma t = T}^{t \geqslant i} P_{a+t}\ p\,(a + i/t - i)\ p\,(a/i)} $$

Sostituendo alle probabilità $p\,(a + i\,/\,t - i)$ le loro espressioni fornite dalla (32) e invertendo a denominatore l'ordine di sommazione si ottiene:

$$ {}_aP_{b_1 \ldots b_k} = \frac{\sum_{\Sigma t = T}^{t \geqslant b} P_{a+t}\ \dfrac{\binom{T-H}{t_1 - b_1 \ldots t_k - b_k}\ \binom{H}{b_1 \ldots b_k}}{\binom{N+T}{a_1 + t_1 \ldots a_k + t_k}}}{\sum_{\Sigma t = T} P_{a+t}\ \dfrac{\sum_{\Sigma i = H}^{i \leqslant t} \binom{T-H}{t_1 - i_1 \ldots t_k - i_k}\ \binom{H}{i_1 \ldots i_k}}{\binom{N+T}{a_1 + t_1 \ldots a_k + t_k}}} $$

ma [23]:

$$ (34) \qquad \sum_{\Sigma i = H}^{i \leq t} \binom{T-H}{t_1 - i_1 \ldots t_k - i_k}\ \binom{H}{i_1 \ldots i_k} = \binom{T}{t_1 \ldots t_k} $$

---

(23) La formula (34) si potrebbe ritener nota, tuttavia ne darò qui una rapida dimostrazione ottenuta imitando un ragionamento svolto da L. Galvani in appendice alla memoria di C. Gini, cit. (4). Si parta dalle identità:

$$ (d_1 + \ldots + d_k)^H = \sum_{\Sigma i = H} \binom{H}{i_1 \ldots i_k} d_1^{i_1} \ldots d_k^{i_k} $$

$$ (d_1 + \ldots + d_k)^{T-H} = \sum_{\Sigma (t-i) = T-H} \binom{T-H}{t_1 - i_1 \ldots t_k - i_k} d_1^{t_1 - i_1} \ldots d_k^{t_k - i_k} $$

$$ (d_1 + \ldots + d_k)^T = \sum_{\Sigma t = T} \binom{T}{t_1 \ldots t_k} d_1^{t_1} \ldots d_k^{t_k} $$

e si osservi che il prodotto delle prime due è identico alla terza. Ne

e quindi:

$$(35)\qquad {}_aP_{b_1\ldots b_k} = \binom{H}{b_1\ldots b_k} \frac{\sum_{\Sigma t=T}^{t\geqslant b} \frac{\binom{T-H}{t_1-b_1\,.\,.\,t_k-b_k}}{\binom{N+T}{a_1+t_1\ldots a_k+r_k}} P_{a+t}}{\sum_{\Sigma t=H} \frac{\binom{T}{t_1\,.\,.\,t_k}}{\binom{N+T}{a_1+t_1\ldots a_k+t_k}} P_{a+t}}$$

Anche la formula (35) non è al solito senz'altro utilizzabile perchè in generale non si può dire di conoscere le probabilità P_{a+t} che in essa compaiono senza prima conoscere la funzione $\omega(p_1\ldots p_{k-1})$. Ma anche questa volta si può immaginare di aver fatto sulla *v. c. di* LAPLACE un gran numero Z di esperienze, ciascuna composta da $N+T$ esperimenti semplici, ottenendo così delle frequenze (relative) f_{a+t} dei risultati $(a_1+t_1, a_2+t_2, \ldots, a_k+t_k)$. Ciò posto *si ammetta che le frequenze così ottenute forniscano una buona approssimazione delle probabilità incognite,* dopo di che si può scrivere:

$$(36)\qquad {}_aP_{b_1\ldots b_k} = \binom{H}{b_1\,.\,.\,b_k} \frac{\sum_{\Sigma t=T}^{t\geqslant b} \frac{\binom{T-H}{t_1-b_1\ldots t_k-b_k}}{\binom{N+T}{a_1+t_1\ldots a_k+t_k}} f_{a+t}}{\sum_{\Sigma t=T} \frac{\binom{T}{t_1\ldots t_k}}{\binom{N+T}{a_1+t_1\ldots a_k+t_k}} f_{a+t}}$$

---

segue:

$$(d_1+\ldots+d_k)^H(d_1+\ldots+d_k)^{T-H} = \sum_{\Sigma t=T}\left[\sum_{\Sigma i=H}^{i\leqslant t}\binom{H}{i_1\,.\,.\,i_k}\binom{T-H}{t_1-i_1\,.\,.\,t_k-i_k}\right] d_1^{t_1}\ldots d_k^{t_k}$$

ed identificando il secondo membro di quest'ultima formula con quello della terza delle precedenti identità si ottiene la relazione che si voleva stabilire.

Questa è la formula che estende al caso di quanti si vogliono eventi necessari e incompatibili il metodo, di C. Gini, dei risultati indiretti.

7. – *Esempi.* — In questo ultimo numero mi propongo di chiarire, sviluppando pratici esempi suggeritimi dal Prof. Gini (basati però su dati puramente fittizi), il contenuto del metodo dei risultati, diretti o indiretti, e l'uso delle formule (30) e (36) relative a tale metodo.

Esempio 1°. — È noto che i colombi fanno normalmente covate di due uova, di modo che ogni covata, giunta *regolarmente* a termine, e solo di queste mi occuperò nel seguito, dà luogo, riguardo al sesso dei nati, ad uno dei tre eventi necessari e incompatibili:

E_1 = un maschio e una femmina
E_2 = due maschi
E_3 = due femmine

Si supponga ora che in una certa stazione per l'allevamento dei colombi si siano registrati i risultati ottenuti in dieci covate da parecchie coppie, risultati raccolti coppia per coppia.

Si sappia ora che una coppia di colombi in otto covate precedenti ha avuto quattro volte un maschio e una femmina, tre volte due maschi e una volta due femmine. Con il nostro simbolismo diremo che la coppia ha presentato il risultato (4, 3, 1)

Si vogliono informazioni sulle covate future di questa coppia; voler rispondere con precisione a tale quesito è naturalmente un'impresa che non ha senso, bisogna limitarsi a precisare come si distribuiscono i risultati futuri di tutte quelle coppie che hanno presentato il risultato (4, 3, 1). A tale quesito si può rispondere con le formule (30) e (36) approfittando dei risultati già ottenuti in 10 covate da un gran numero di coppie, risultati che abbiamo già supposto di conoscere.

È chiaro che nel nostro caso interessano solamente i seguenti sei risultati:

(6, 3, 1), (5, 4, 1), (4, 5, 1), (5, 3, 2), (4, 4, 2), (4, 3, 3)

di cui quindi, con i dati a disposizione, occorrerebbe calcolare le frequenze relative. Ma, data la natura delle formule che vogliamo

applicare, non è necessario conoscere proprio i valori di tali frequenze, ma solo valori ad esse proporzionali. Supponiamo pertanto che si abbiano le proporzioni:

$$f_{631} : f_{541} : f_{451} : f_{532} : f_{442} : f_{433} = 6 : 5 : 4 : 2 : 2 : 1.$$

Approfittando della formula (36), cioè del metodo dei risultati indiretti, possiamo calcolare la probabilità ${}_{431}P_{100}$ che in una successiva covata si ottengano un maschio e una femmina, oppure la probabilità ${}_{431}P_{010}$ che in una successiva covata si ottengano due maschi, oppure la probabilità ${}_{431}P_{001}$ che in una successiva covata si ottengano due femmine.

In base ai dati che supponiamo di conoscere si ottiene:

$$\begin{aligned} {}_{431}P_{100} &\simeq 0{,}5626 \\ {}_{431}P_{010} &\simeq 0{,}3600 \\ {}_{431}P_{001} &\simeq 0{,}0774 \\ \text{Totale} \quad &= 1{,}0000 \end{aligned}$$

Approfittando invece della formula (30) cioè del metodo dei risultati diretti, possiamo fare previsioni sui risultati di due covate successive; e nel nostro caso si ottiene:

$$\begin{aligned} {}_{431}P_{200} &\simeq 0{,}3425 \\ {}_{431}P_{110} &\simeq 0{,}3670 \\ {}_{431}P_{020} &\simeq 0{,}1468 \\ {}_{431}P_{101} &\simeq 0{,}0733 \\ {}_{431}P_{011} &\simeq 0{,}0594 \\ {}_{431}P_{002} &\simeq 0{,}0110 \\ \text{Totale} \quad &= 1{,}0000 \end{aligned}$$

Esempio 2°. — In un certo ufficio lavorano tre dattilografe, due delle quali una mattina telefonano che per quel giorno non verranno in ufficio; il direttore deve allora riporre tutte le sue speranze sulla terza dattilografa ed ha interesse ad aver subito previsioni su che cosa essa farà. Sono a priori possibili tre casi:

E_1 = non viene affatto
E_2 = viene in mattinata
E_3 = non viene in mattinata, ma si presenta il pomeriggio.

Egli non potrà certo prevedere quello che succederà quel giorno, ma, disponendo di un registro di presenze potrà stabi-

lire come si distribuiscono i tre eventi nei giorni in cui fin dall'inizio si presenta il risultato (2, 0, 0).

In questo caso i risultati che interessano sono solo i seguenti tre: (3, 0, 0), (2, 1, 0), (2, 0, 1); e per le rispettive frequenze, dal registro delle presenze, si ricavi:

$$f_{300} : f_{210} : f_{201} = 1 : 6 : 3$$

Applicando la formula (30) egli allora otterrà:

$$\begin{aligned} {}_{200}P_{100} &\simeq 0{,}25 \\ {}_{200}P_{001} &\simeq 0{,}50 \\ {}_{200}P_{001} &\simeq 0{,}25 \\ \text{Totale} \quad &= 1{,}00 \end{aligned}$$

Il direttore di quell'ufficio può ora tranquillizzarsi perché ha circa il 75 % di probabilità che la terza dattilografa venga in ufficio magari solo il pomeriggio.

Ben diverso sarebbe invece stato il suo atteggiamento se dal registro delle presenze avesse ricavato:

$$f_{300} : f_{210} : f_{201} = 7 : 6 : 3$$

perchè in tal caso avrebbe trovato:

$$\begin{aligned} {}_{200}P_{100} &\simeq 0{,}7 \\ {}_{200}P_{010} &\simeq 0{,}2 \\ {}_{200}P_{001} &\simeq 0{,}1 \\ \text{Totale} \quad &= 1{,}0 \end{aligned}$$

Naturalmente perchè questi due esempi abbiano senso bisogna supporre che si verifichino tutte quelle ipotesi che a suo tempo abbiamo fatto durante lo svolgimento della teoria.

Roma, dicembre 1947.

FRED C. LEONE

Some extensions of Bayes' theorem

1. Introduction

The present paper seeks to generalize Bayes' theorem in two respects and to consider certain other related problems. Many controversies have raged over this theorem since the first posthumous publication of the work. M. G. Kendall [1] in presenting an exposition of the various theories of probability states that "there is nothing necessarily incompatible in the varying views which are currently held ; ... the authorities are either saying the same thing in different ways or can only disagree because of avoidable latent differences in their premises or their field of discussion ». Most objections raised against Bayes' theorem are not on the mathematics of the theorem, but on the propriety of making the assumption that « complete prior ignorance » permits or justifies a rectangular distribution of a priori probability. Although such men as Karl Pearson and Edgeworth have attempted to justify this assumption, we shall not enter the arena.

The need for and potential use of such inferences as Bayes' theorem provides arise in attempts to make use of past inspection data from a supplier of industrial product. Here it is often the case that a consumer has a record of the results of 100% inspection of a series of lots from a supplier. He wants to make a decision on the basis of a sample. Now if the quality level of a supplier has consistently been highly acceptable except on rare occasions, then a small sample yielding « good » results will provide all the evidence needed. But another supplier with a highly

(1) M. G. Kendall, « *On the reconciliation of the theories of probability* », Biometrika, Vo. 36, pp. 101-116 (1949).

variable lot quality may need a more sizable sample per lot to provide an adequate a posteriori probability of the lot being a satisfactory one.

Now with the cooperation of a number of industries it has been found that there are such records available in receiving inspection in plants, that a priori distributions of lot fraction defectives do take quite a variety of distributions, many being non-normal, such as J-shaped and U-shaped, and that at present the only use being made of such records is in « educated » guesses by the chief inspectors as to how to proceed. Such an industrial situation and problem gives practicality to the theory here presented. It is reasonable to suppose that the results will prove useful outside of the field of incoming inspection, if assumptions are carefully checked.

To lay precise groundwork let us first consider the following interpretation of Bayes' theorem and then generalize it. Consider a large scale experiment in which we have say 1000 beads in each of 101 boxes, but the proportion of red beads is .00 in the first, .01 in the second, and up to 1.00 in the last. We choose a box at random (equal probability) and then from the chosen box choose ten beads with replacement. We decide to concern ourselves only with those experiments resulting in precisely one red bead in the sample. The overall experiment is repeated a huge number of times, and we keep a record of the number of times each box was the medium through which the one-red-in-ten samples arose. The relative frequencies of the several boxes do not form a rectangular distribution but tend to approach an « a posteriori » distribution given by Bayes' theorem. Thus the a posteriori probability of the box being of proportion p red is

$$P_p = \frac{\frac{1}{101} C(10,1)\, p\, (1-p)^9}{\sum_{p=.00}^{10.0} \frac{1}{101} C(10,1)\, p\, (1-p)^9}.$$

The generalization is obvious. First we no longer require a uniform a priori distribution of box or lot proportions but instead consider any kind of distribution actually or potentially arising. Since lots are of finite size, this would require a discrete function of a priori lot quality. We further generalize this to permit a

continuous function to be used. When treated and applied as in the foregoing illustration, then it is perfectly reasonable to work with the a posteriori distribution of lot proportions like other ordinary distributions. Thus we might find areas such as the a posteriori probability of the lot quality being less than a certain amount. We can also make use of the moments of the a posteriori distribution of lot quality.

All of this does appear to make practical sense in receiving inspection if (1) there is enough past information on a supplier's lot quality of a product to at least approximate the a priori probability distribution and if (2) there is not much serial correlation in lot quality from the supplier so that we can assume that we are taking a random sample lot from the a priori distribution. To take account of serial correlation in deciding on a lot (that is, using the previous lot or two in addition to the a priori distribution) is an interesting theoretical problem not dealt with in this paper.

The problem will now be developed as a theoretical one of interest, followed by a few examples. We shall assume various theoretical distributions as the a priori probability distributions and from each of these obtain an a posteriori distribution (1).

(1) After writing this paper I have learned of the works of Professor Corrado Gini on this subject (see N° 4, 5, 6, 7, 8 of the bibliography). I have thus learned that since 1911 (4) Gini had considered the hypothesis of a distribution of *a priori* probability of Pearson's Type I, and had drawn therefrom the formulas for the *a posteriori* probability, calling attention to the essential difference of the results thus obtained as compared to those obtained on the basis of the traditional hypothesis of a rectangular distribution of the *a priori* probability. Returning to this matter in 1943 (6,7) Gini insisted on the inaccuracy of the statement that in a great number of observations the distribution of a *priori* probability is practically a matter of indifference for the determination of the *a posteriori* probability. He has moreover (7) illustrated the method by which one can calculate the parameters of the distribution of *a priori* probability of Pearson's Type I, availing oneself of the knowledge already acquired in the past on the phenomenon under consideration, and he has applied it to the case of the testing of lots by means of samples, showing the decisive importance the values of these parameters may have for the results of the test.

The use of the knowledge acquired in the past is the central idea of the 1911 paper (4). In this Gini stated – and in 1946 he recalled and developed (8) – the *method* which he calls that of *results* (direct and indirect) which in some cases allows of a rational and complete use of such knowledge. This method, propounded and applied by Gini characters of two modalities, has been generalised to meet the case of characters of more modalities by G. Pompilj (16, 17).

2. A Priori and A Posteriori Distributions

2.1 Let us assume an a priori distribution, $f(p)$. This might be a superpopulation of lots each of which is characterized by the parameter p. Then the a posteriori distribution of p for a randolmy chosen lot can be represented as

$$(1) \qquad F(p \mid n,c)\,dp = \frac{f(p)\,C(n,c)\,p^c(1-p)^{n-c}\,dp}{\int_a^b f(p)\,C(n,c)\,p^c(1-p)^{n-c}\,dp}$$

where c defectives are found in a sample of n and (a, b) are the limits of p. The a posteriori probability that p lies between α and β for this particular lot is

$$(2) \qquad F(\alpha \leqq p \leqq \beta \mid n,c) = \frac{\int_\alpha^\beta f(p)\,p^c(1-p)^{n-c}\,dp}{\int_a^b f(p)\,p^c(1-p)^{n-c}\,dp}$$

where

$$a \leqq \alpha < \beta \leqq b.$$

2.2 When the a priori distribution is rectangular we have

$$(3) \qquad f(p)\,dp = \frac{1}{b-a}\,dp$$

where $a \leqq p \leqq b$. The a posteriori distribution of the paramenter is then:

$$(4) \qquad F(p \mid n,c) = B^{-1}_{(b-a)}(c+1,\ n-c+1)\,p^c(1-p)^{n-c}$$

where we define $B^{-1}_{(b-a)}$ as the reciprocal of the difference of the incomplete beta functions, namely $(B_b - B_a)^{-1}$. In other words, given the a priori rectangular distribution of lots, we consider

the theoretical distribution of all samples of n producing exactly c defectives. Then the probability density function (or relative a posteriori probability) that a particular sample came from a lot of fraction defective equal to p can be found from equation (4). The k^{th} moment about the origin of $F(p \mid n, c)$ is

$$\nu_{k:p} = \frac{B_{(b-a)}(c+k+1, n-c+1)}{B_{(b-a)}(c+1, n-c+1)} . \tag{5}$$

If $a = 0$, then equations (4) and (5) reduce to

$$F(p \mid n, c) = B_b^{-1}(c+1, n-c+1)\, p^c (1-p)^{n-c} \tag{6a}$$

and

$$\nu_{k:p} \frac{B_b(c+k+1, n-c+1)}{B_b(c+1, n-c+1)} . \tag{6b}$$

If, moreover, $b = 1$ and therefore $0 \leqq p \leqq 1$, equations (4) and (5) become

$$F(p \mid n, c) = (n+1)\, C(n, c)\, p^c (1-p)^{n-c} \tag{7a}$$

and

$$\nu_{k:p} = \frac{(n+1)!\,(c+k)!}{(n+k+1)!\,c!} \tag{7b}$$

where $C(n, c)$ is a binomial coefficient. Equation (7a) is a continuous binomial distribution of p with the constant coefficient such that

$$\int_0^1 F(p \mid n, c)\, dp = 1 .$$

2.3 Consider the situation where the a priori distribution of lots is trangular. Such a case is a first approximation to some practical conditions. If this distribution is

$$f(p) = K(a - p) \tag{8}$$

with $0 \leqq p \leqq a \leqq 1$ and $K = 2/a^2$, the a posteriori distribution of p is

$$F(p \mid n, c) = \frac{(a - p)\, p^c (1 - p)^{n-c}}{a\, B_a(c+1, n-c+1) - B_a(c+2, n-c+1)}. \tag{9}$$

2.4 *A* distribution which can readily occur in practice is the binomial distribution. For example, we may have a binomial distribution of lot fraction defective where the lots arise from an industrial process with constant fraction defective. The *a* priori probability of x defectives in a lot of size N is then

$$f(x) = C(N, x)\, p^x (1 - p)^{N-x} \tag{10}$$

where p defines the process average. If c defectives are found in a random sample of n, the a posteriori probability that the fraction defective is x/N is

$$P(x/N \mid n, c) = \tag{11}$$

$$\frac{C(N, x)\, p^x (1 - p)^{N-x}\, C(n, c)\, (x/N)^c (1 - x/N)^{n-c}}{\sum_{x=0}^{N} C(N, x)\, p^x (1 - p)^{N-x}\, C(n, c)\, (x/N)^c (1 - x/N)^{n-c}}.$$

By expanding $(1 - x/N)^{n-c}$ and simplifying, equation (11) becomes

$$P(x/N \mid n, c) = \frac{C(N, x)\, p^x (1 - p)^{N-x}\, x^c (N - x)^{n-c}}{\sum_{i=0}^{n-c} (-1)^i\, C(n - c, i)\, N^{n-c-i}\, \nu_{c+i:x}} \tag{12}$$

where ν_k is the k^{th} moment about the origin of the a priori binomial distribution.

2.5 Next suppose the a priori distribution function of the superpopulation of lots is normal, namely

$$f(p) = \frac{1}{\sqrt{2\pi}\,\sigma_p} e^{-\frac{(p-v_{1:p})^2}{2\sigma_p^2}} . \tag{13}$$

The natural limits for a normal function are, of course, $-\infty$ to ∞, whereas the limits of p are 0 to 1. Hence we must consider 0 and 1 as the practical limits of the function. This is not likely to be serious if the limits are are at least $2\sigma_p$ from $V_{1:p}$. The a posteriori distribution becomes

$$F(p \mid n, c) = \frac{\frac{1}{\sqrt{2\pi}\,\sigma_p} e^{-\frac{(p-v_{1:p})^2}{2\sigma_p^2}} p^c (1-p)^{n-c}}{\int_0^1 \frac{1}{\sqrt{2\pi}\,\sigma_p} e^{-\frac{(p-v_{1:p})^2}{2\sigma_p^2}} p^c (1-p)^{n-c}\, dp} . \tag{14}$$

If we let $t = \frac{p - v_{1:p}}{\sigma_p}$ we get

$$G(t \mid n,c) = \frac{e^{-\frac{t^2}{2}} (v_{1:p} + \sigma_p t)^c (1 - v_{1:p} - \sigma_p t)^{n-c}}{\int_{-\frac{v_{1:p}}{\sigma_p}}^{\frac{1-v_{1:p}}{\sigma_p}} e^{-\frac{t^2}{2}} (v_{1:p} + \sigma_p t)^c (1 - v_{1:p} - \sigma_p t)^{n-c}\, dt.} \tag{15}$$

By the use of the binomial expansion and incomplete gamma functions, equation (15) simplifies to

$$G(t \mid n, c) = \frac{e^{-\frac{t^2}{2}} (v_{1:p} + \sigma_p t)^c (1 - v_{1:p} - \sigma_p t)^{n-c}}{\sum_{i=0}^{c} \sum_{j=0}^{n-c} C(c,i)(-1)^j C(n-c,j)\, v_p^{c-i} (1 - v_{1:p})^{n-c-j} \sigma_p^{j+i}}$$

$$2^{-\frac{i+j-1}{2}} \left[\Gamma_{\frac{(1-v_{1:p})^2}{2\sigma_p^2}} \left(\frac{i+j+1}{2} \right) + (-1)^{i+j} \Gamma_{\frac{v_{1:p}^2}{2\sigma_p^2}} \left(\frac{i+j+1}{2} \right) \right]. \tag{16}$$

2.6 If the a priori distribution is a Pearson Type III, that is,

$$g(t) = K(A + t)^{A2-1} e^{-At} \tag{17}$$

where

$$K = \frac{e^{-A^2} A^{A^2}}{\Gamma(A^2)}, \quad A = \frac{2}{\alpha_3} \quad \text{and} \quad t = \frac{p - \nu_{1:p}}{\sigma_p}$$

the a posteriori distribution of t for a randomly chosen lot is then

$$G(t \mid n, c) = \frac{K(A+t)^{A2-1} e^{-At} (\nu_{1:p} + \sigma_p t)^c (1 - \nu_{1:p} - \sigma_p t)^{n-c}}{K \int_{\frac{-\nu_{1:p}}{\sigma_p}}^{\frac{1-\nu_{1:p}}{\sigma_p}} (A+t)^{A2-1} e^{-At} \left[\sum_{i=0}^{c} C(c, i)\, \nu_{1:p}^{c-i} (\sigma_p t)^i\right] \times \left[\sum_{j=o}^{n-c} (-1)^j C(n-c, j) (1 - V_{1:p})^{n-c-j} (\sigma_p t)^j\right] dt} \tag{18}$$

In order to determine the value of the denominator of equation (18), let us make the following approximation :

$$\int_{\frac{-\nu_{1:p}}{\sigma_p}}^{\frac{1-\nu_{1:p}}{\sigma_p}} \doteq \int_{-A}^{\infty}.$$

If the lower limit of t, that is, $\frac{-\nu_{1:p}}{\sigma_p}$ is less than the natural lower limit of the Pearson Type III, that is, $-A$, then the area from $-\nu_{1:p}/\sigma_p$ to some upper limit is equal to the area from $-A$ to this same upper limit. Further, the relative error in using ∞ as the upper limit in place of $\frac{1 - \nu_{1:p}}{\sigma_p}$ is negligible. For the values $\nu_{1:p} = 0.12$, $\sigma_p = 0.04$, $\alpha_3 = 1$, $c = 2$ and $n = 25$, the relative error, which is approximately

$$\int_{\frac{1-\nu_{1:p}}{\sigma_p}}^{\infty} \div \left(\int_{-A}^{\infty} \quad \int_{\frac{1-\nu_{1:p}}{\sigma_p}}^{\infty} \right)$$

is of the magnitude 10^{-11}.

Hence we have as an approximation

$$G(t \mid n, c) =$$

(19)

$$\frac{K(A+t)^{A^2-1} e^{-At} (\nu_{1:p} + \sigma_p t)^c (1 - \nu_{1:p} - \sigma_p t)^{n-c}}{\sum_{i=0}^{c} \sum_{j=0}^{n-c} C(c, i) (-1)^j C(n-c, j) \nu_{1:p}^{c-i} (1 - \nu_{1:p})^{n-c-i} \sigma_p^{i+j} \alpha_{i+j}}$$

where α_{i+j} is the $i+j$ standard moment of $f(p)$ and $g(t)$.

2.7 Consider finally the a priori distribution as Pearson's Type I. Hence

$$f(p) = \frac{K}{\sigma_p} \left(\frac{p - \nu_{1:p}}{\sigma_p} - r_1 \right)^{m_1} \left(r_2 - \frac{p - \nu_{1:p}}{\sigma_p} \right)^{m_2} \tag{20}$$

or in standard units

$$g(t) = K (t - r_1)^{m_1} (r_2 - t)^{m_2} \tag{20a}$$

where

$$K = [(r_2 - r_1)^{m_1 + m_2 + 1} B(m_1 + 1, m_2 + 1)]^{-1}.$$

The parameters r_1, r_2, m_1 and m_2 are those given by C. C. Craig in his paper on the Pearson System of Curves. (1)

The a posteriori distribution of t from the randomly chosen lot is then

$$G(t \mid n, c) =$$

(21)

$$\frac{K (t - r_1)^{m_1} (r_2 - t)^{m_2} (\nu_{1:p} + \sigma_p t)^c (1 - \nu_{1:p} - \sigma_p t)^{n-c}}{\sum_{i=0}^{c} \sum_{j=0}^{n-c} C(c, i) (-1)^j C(n-c, j) \nu_{1:p}^{c-i} (1 - \nu_{1:p})^{n-c-j} \sigma_p^{i+j} \alpha_{i+j}}.$$

(1) C. C. Craig, « *A new exposition and chart for the Pearson system of frequency curves*, Annals of Matematical Statistics, Vol. 7 pp. 16-28 (1936).

3. Bayes' Theorem Applied to Sampling Inspection

One of the chief contributions which can be made by Bayes' theorem is its application to sampling inspection. With this we can gain new information about a specific lot from which a random sample is taken provided we have some past data on the distribution of lots from which this lot has come. Moreover, we can test the effectiveness of our sampling plan and consequently tighten the plan, retain or even loosen it.

Suppose p_2 is the fraction defective of a marginally poor lot. Consider A as the a posteriori probability that $0 \leqq p \leqq p_2$ if the lot is accepted. Then

$$A = \frac{\int_0^{p_2} P\,(acc \mid p)\; P\,(p)\; dp}{\int_0^1 P\,(acc \mid p)\; P\,(p)\; dp} \tag{22}$$

where $P(p)$ is the a priori distribution and $P(acc \mid p)$ is the probability of acceptance if the fraction defective of the lot is p. $P(acc \mid p)$ can be obtained from the operating characteristic curve for the sampling plan in effect.

Equation (22) becomes

$$A = \frac{\sum_{c=0}^{a} \int_0^{p_2} p^c\,(1-p)^{n-c}\, f\,(p)\; dp}{\sum_{c=0}^{a} \int_0^1 p^c\,(1-p)^{n-c}\, f\,(p)\; dp} \tag{23}$$

where $f\,(p)$ is the a priori distribution of lots and a is the acceptance number, that is, the lot is accepted if $c \leqq a$. By this approach we can get the a posteriori distribution of fraction defective for passed lots.

Finally, we may proceed as follows:

(24)
$$\begin{cases} \text{If } 0.95 < A < 0.99, \text{ retain the present plan.} \\ \text{If } 0.99 \leqq A, \text{ change the plan to a looser level.} \\ \text{If } A \leqq 0.95, \text{ tighten the sampling plan or perhaps use 100\% inspection.} \end{cases}$$

4. Distributions of Lot Quality

4.1 Let us consider two examples of distributions of lot quality which are taken from industrial data. These distributions concern populations of lots each of which is characterized by its fraction defective. In each example, after the frequency polygon is obtained from the raw data, a theoretical a priori probability distribution is sought to fit these data and the various parameters are obtained. The chi-square test is then used to test the goodness of fit. Finally, by making use of the methods shown in the preceding section, the a posteriori probabilities can be obtained.

Whenever the probability distribution does not afford a satisfactory fit, an empirical problem can be set up. Hence, for a sample of size n from a random lot in which c defectives are found, the a posteriori probability that p' [1] $= p_1$ is given by the formula

$$(25) \qquad P(p' = p_1 \mid n, c) = \frac{P(p' = p_1)\, p'^c (1 - p')^{n-c}}{\sum_{p'=0}^{1} P(p' = p_1)\, p'^c (1 - p')^{n-c}}$$

where $P(p' = p_1)$ is the relative frequency of p_1 as taken from the frequency polygon.

4.2 For the first example let us take a case of incoming inspection. A specific product to be used by an industry is supplied by a number of vendors. Over a period of months this product has been 100% inspected at incoming inspection and the quality

---

[1] In Statistical Quality Control p' is used rather than p to designate the fraction defective of a lot.

is considered « good ». The population of lots to be examined is made up of 92 lots in all. These lots vary in size since this factor is not controllable at incoming inspection. The data are then tabulated and the frequency polygon is obtained. This is shown in Figure 1. The parameters of the a priori probability distribution are then

$$(26)\qquad \left\{\begin{array}{ll} V_{1;p'} & = 0.313913 \\ \sigma_{p'} & = 0.368684 \\ \alpha_3 & = 0.8332 \\ \alpha_1 & = 2.1162 \\ \delta & = -\ 0.7525 \end{array}\right.$$

Using the values of α_3 and δ and referring to Craig's chart of the Pearson System [1], we find that the theoretical distribution of the Pearson System is Type I (U-shaped). The equation of this distribution is

$$(27)\qquad g(t) = K\,(t - r_1)^{m_1}\,(r_2 - t)^{m_2}$$

where

$$K = [(r_2 - r_1)^{m_1+m_2+1}\,B\,(m_1 + 1\,,\, m_2 + 1)]^{-1}$$

and

$$t = \frac{p' - v_{1:p'}}{\sigma_{p'}}$$

Solving for the parameters r_1, r_2, m_1 and m_2, we obtain as the theoretical distribution

$$(28)\qquad g(t) = 0.2188\,(t + 0.8479)^{-0.8011}\,(1.9550 - t)^{-0.5415}.$$

---

[1] See note 1 of chaptre 2.

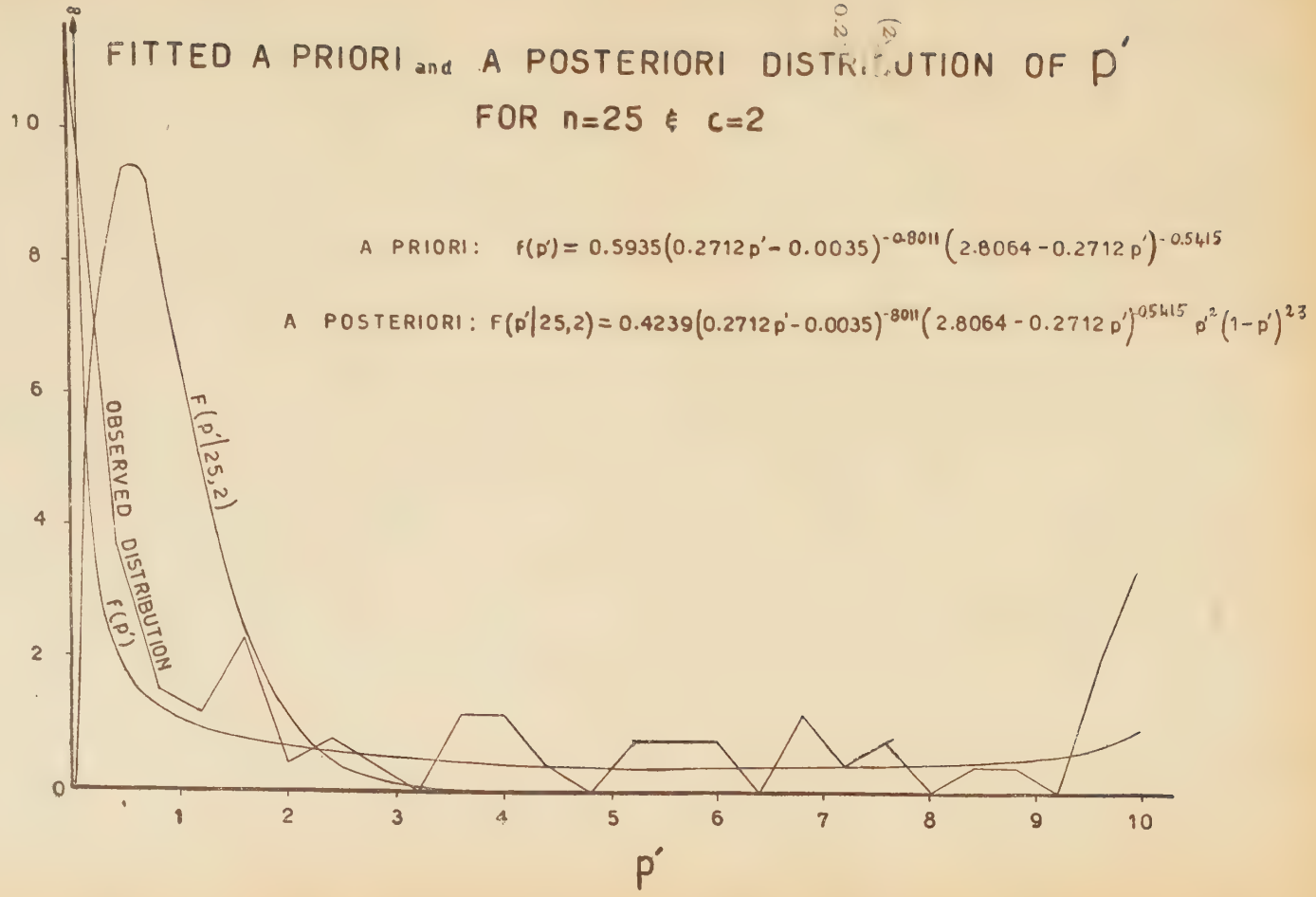

FITTED A PRIORI and A POSTERIORI DISTRIBUTION OF p′
FOR n=25 & c=2
A PRIORI: f(p′) = 0.5935(0.2712 p′ − 0.0035)^-0.8011 (2.8064 − 0.2712 p′)^-0.5415
A POSTERIORI: F(p′|25,2) = 0.4239(0.2712 p′ − 0.0035)^-8011 (2.8064 − 0.2712 p′)^-05415 p′^2 (1−p′)^23
F(p′|25,2)
OBSERVED DISTRIBUTION
f(p′)
0
2
4
6
8
10
1
2
3
4
5
6
7
8
9
10
p′

Fig. 1

The chi-square test yelds a value of 2.574 for chi-square. The probability of chi-square equal to or greater than 2.574 for 5 degrees of freedom is 0.76. That is to say, the probability of as bad or worse a fit is 0.76 by chance alone. Hence the observed data could perfectly well have come from such a type of population.

Now let us suppose that equation (28) describes the theoretical dustribution of the lots, and a sample of 25 is taken and that 2 pieces are found to be defective. The a posteriori distribution is then

$$G(t \mid 25,2) =$$

(29)

$$0.2188\,(t+0.8479)^{-0.8011}\,(1.9550-t)^{-0.5415}\,(\sigma_{p'}t+\nu_{1:p'})^2\,(1-\sigma_{p'}t-\nu_{1:p'})^{23}$$
$$\sum_{i=0}^{2}\sum_{j=0}^{23} C(2,i)\,(-1)^i\,C(23,j)\,(\nu_{1:p'})^{2-i}\,(1-\nu_{1:p'})^{23-j}\,\sigma_{p'}{}^{i+j}\alpha_{i+j}$$

$$= 1563\,(t+0.8479)^{-0.8011}\,(1.9550-t)^{-0.5415}\,(0.3687\,t+$$

(30)

$$0.3139)^2\,(0.6861-0.3687\,t)^{23}\,.$$

The α's in formula (29) are derived from the original data and the recursion formula

$$\alpha_k = \frac{k-1}{2-(k-3)\,\delta}\left[(2+\delta)\,\alpha_{k-2}+\alpha_3\,\alpha_{k-1}\right].$$

The a priori distribution $f(p')$ and the a posteriori distribution $F(p' \mid n,c)$ for $n=25$, $c=2$ are also included in Figure 1.

4.3 Let us now consider a bell-shaped distribution. In the manufacture of a product plant X makes various component parts before assembling them at a central area. At this stage one of the parts must be 100% inspected while the others may be sampled. Over a period of seven months 158 lots have been 100% inspected. The size of these lots varies, but the average lot size is about 3000 pieces with none of the lots under 2000 pieces. We wish to determine the a priori probability distribution of lots. We also wish to obtain the a posteriori probability distribution. That is to say, upon chosing a random lot and taking a sample of size n with c defectives, we want the probability, a

posteriori, that the fraction defective of this lot is equal to p' or is between two values of p', say p_1 and p_2.

The values of the parameters of the a priori distribution are

$$(31) \qquad \begin{cases} \nu_{1:p'} = 0.049304 \\ \sigma_{p'} = 0.028994 \\ \alpha_3 = 1.3294 \\ \alpha_4 = 5.6162 \\ \delta = -0.00868 \end{cases}$$

These parameters suggest a Pearson Type I (bell-shaped) distribution. However'since δ is so close to zero, Pearson's Type III may be tried as the theoretical distribution for the data. A general Pearson Type III is

$$(32) \qquad g(t) = K(A+t)^{A^2-1} e^{-At}$$

where

$$A = \frac{2}{\alpha_3}, \quad t = \frac{p' - \nu_{1:p'}}{\sigma_{p'}}$$

and

$$K = \frac{e^{-A^2} A^{A^2}}{\Gamma(A^2)}.$$

Substituting the values of the parameters in equation (32), we have

$$(33) \qquad g(t) = 0.2297\ (1.5042 + t)^{1.2626}\ e^{-1.5042t}.$$

The chi-square test yelds a probability of 0.79 ; that is to say, the probability of at least as bad a fit is 0.79 by chance alone. Hence we may say that the observed data could have come from such a population.

Finally, upon taking a sample of size n equal to 25 from a random lot and finding c equal to 2 defectives, we can obtain the a posteriori distribution. This distribution, in terms of the standard unit t, is

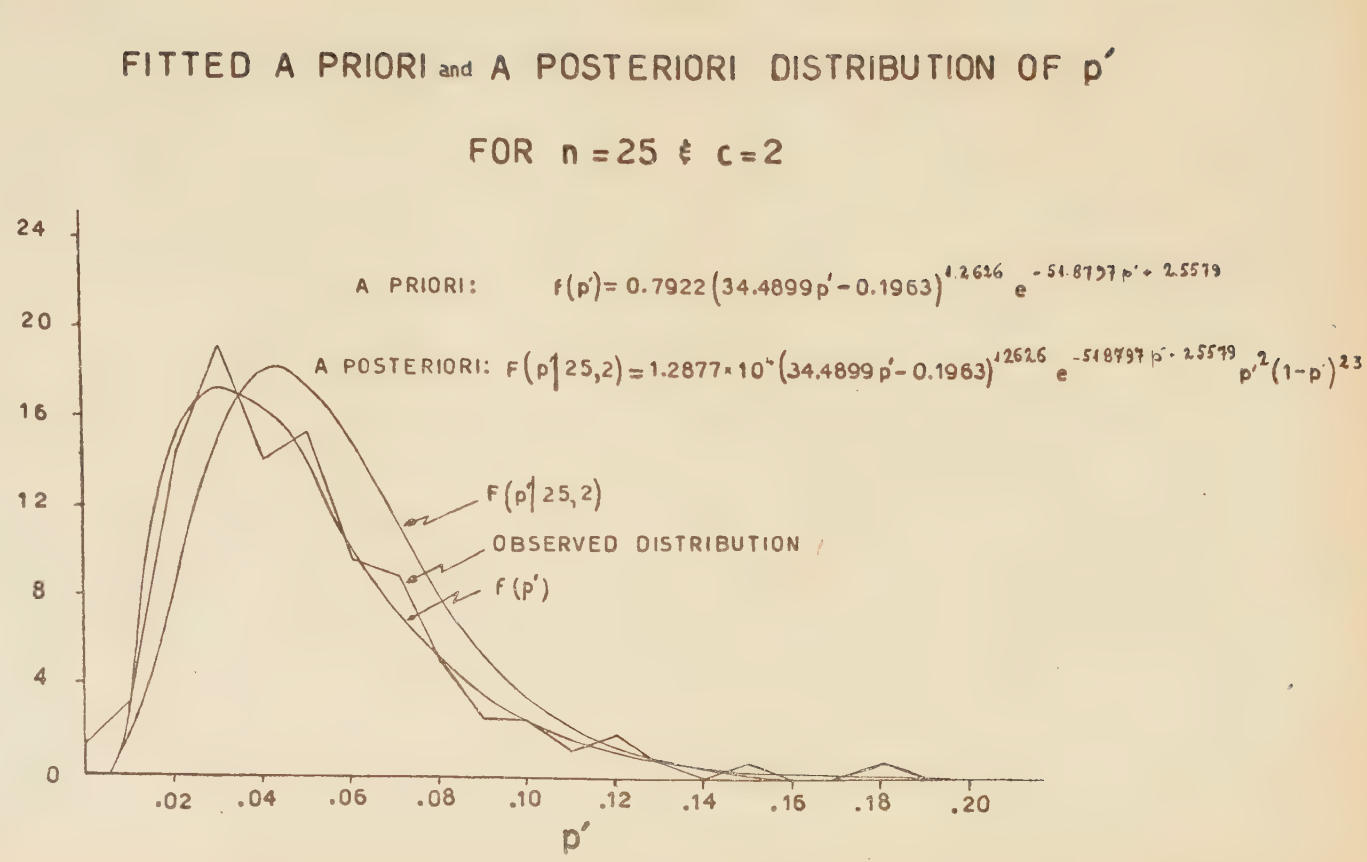

FITTED A PRIORI and A POSTERIORI DISTRIBUTION OF p′
FOR n = 25 & c = 2
A PRIORI: f(p′) = 0.7922 (34.4899 p′ − 0.1963)^1.2626 e^(−51.8797 p′ + 2.5579)
A POSTERIORI: F(p′|25,2) = 1.2877 × 10^4 (34.4899 p′ − 0.1963)^1.2626 e^(−51.8797 p′ + 2.5579) p′^2 (1−p′)^23
F(p′|25,2)
OBSERVED DISTRIBUTION
f(p′)
24
20
16
12
8
4
0
.02
.04
.06
.08
.10
.12
.14
.16
.18
.20
p′

Fig. 2

$$G(t \mid 25,2) = 373.35\,(1.5042 + t)^{1.2626}\, e^{-1.5042\,t} \times$$

(34)

$$(0.028994\,t + 0.049304)^2\,(0.950696 - 0.028994\,t)^{23}.$$

Figure 2 presents a graphical representation of the above example. The frequency polygon represents the observed data. Each original ordinate was multiplied by a constant factor in order to make the area under this polygon equal to unity. The theoretical a priori probability distribution of p' is given by $f(p')$. This is a Pearson Type III. $F(p' \mid n, c)$ is the a posteriori probability distribution of p'.

5. Conclusion

The application of the theorem of Bayes and its extensions can offer some assistance to many of the statistical problems of an inverse nature. In various fields of application of statistics this theorem can be used advantageously. In sampling inspection, where each sampling plan includes some risk, we can minimize this risk and even reduce the amount of sampling in many cases. This can be done by considering all the past information available and from this obtaining an a posteriori view of our product. This theorem and its extensions then enable us to make use of both past data and the information gained from our present sample in order that we may get as clear a picture as possible.

BIBLIOGRAPHY

1. J. Berkson, *Bayes' thearem*, « Annals of Math » Stat., Vol. 1 (1930), pp. 42-56.

2. C. C. Craig, *A new exposition and chart for the Pearson system of curves*, « Annals of Math Stat »., Vol. 7 (1936), pp. 16-28.

3. W. Edward Deming, ed *Facsimile of Two Papers by Bayes*, Graduate School of the Department of Agriculture, Wash., D.C.

4. C. Gini, *Considerazioni sulle probabilità a posteriori e applicazioni al rapporto dei sessi delle nascite umane*, « Studi Economico-Giuridici della Università di Cagliari », a. III (1911) pp. 77-113, article reprinted in full in « Metron », Vol. XV, (1949), pp. 133-169.

5. C. Gini, *I pericoli della statistica*, « Atti della I riunione scientifica della Società Italiana di Statistica » (1939) pp. 1-44. and « Rivista di Politica Economica », fasc. 11 (1939) pp. 901-24.

6. C. Gini, *Testi di significatività*, « Atti della VII riunione della Società Italiana di Statistica » (1943), pp 241-279.

7. C. Gini e G. Livada, *Sulla probabilità inversa nel caso di grandezze tensive ed in particolare sulla sua applicazione a collaudi per ma a mezzo di campioni*, « Atti della VII riunione scientifica della S cietà Italiana di Statistica », (1943), pp. 300-306.

8. C. Gini, *Intorno alle basi logiche e alla portata gnoseologica del metodo statistico*, « Statistica », a. V-VI (1945-46), pp. 1-38 article translated into Spanish under the title *Sobre las bases logicas del metodo estadistico* « Anales del Instituto de Actuario Espanoles », IV-V, (1946) pp. 23-63 and, with several additions, in Portuguese, under the title *Os fundamentos e o alcance do método estatistico*, « Revista brasilera de estatistica », a. IX Nº 35 (1948), pp. 299-342.

9. R. von Mises, *A modification of Bayes' problem*, « Annals of Math. Stat., » Vol. 9 (1938), pp. 256-259.

10. R. von Mises, *On the correct use of Bayes' formula*, « Annals of Math. Stat. », Vol. 13 (1942), pp. 156-165.

11. R. von Mises, *On the foundations of probability and statistics* « Annals of Math. Stat. », Vol. 12 (1941), pp. 191-205.

12. E. C. Molina, *Bayes theorem-an expository presentation*, « Annals of Math. Stat. », Vol. 2 (1933), pp. 23-37.

13. E. S. Pearson, *Bayes' theorem, examined in the light of experimental sampling*, « Biometrika », Vol. 17, pp. 388-442.

14. K. Pearson, *The fundamental problem of practical statistics*, « Biometrika », Vol. 13 (1921), pp. 1-16.

15. L. E. Simon, *An Engineers' Manual of Statistical Methods*, John Wiley & Sons, Inc., New York, 1941.

16. G. Pompilj, *Complementi di Calcolo delle Probabilità*, Veschi, Roma 1948-49.

17. G. Pompilj, *Sul problema inverso della ricerca statistica*, « Metron », Vol. XVI *Roma, 1951.*

CORRADO GINI

Della relazione tra differenza media con ripetizione e scostamento semplice medio dalla mediana [1]

Indichiamo al solito con ${}^1S_M$ lo scostamento semplice medio dalla mediana M, con Δ_R la differenza media con ripetizione tra gli n termini (n essendo un numero pari), con Δ'_R la differenza media con ripetizione tra gli $\frac{n}{2}$ termini superiori alla mediana e con Δ''_R la differenza media con ripetizione tra gli $\frac{n}{2}$ termini inferiori alla mediana. Tra le n^2 differenze che si possono istituire tra gli n termini della serie, $\frac{n^2}{4}$ riguardano coppie di termini superiori alla mediana, $\frac{n^2}{4}$ coppie di termini inferiori alla mediana ed $\frac{n^2}{2}$ coppie di termini di cui uno è superiore e uno è inferiore alla mediana. Indicando con d la media tra le $\frac{n^2}{2}$ differenze che si possono istituire tra un termine superiore e un termine inferiore alla mediana, potremo quindi scrivere:

$$n^2 \Delta_R = \frac{n^2}{4} \Delta'_R + \frac{n^2}{4} \Delta''_R + \frac{n^2}{2} d$$

[1] Nota presentata alla riunione del 18 giugno 1946 del Seminario di Statistica dell'Istituto di Statistica della Università di Roma.

o anche (1):

$$\Delta_R = \frac{1}{2} d + \frac{1}{4} (\Delta'_R + \Delta''_R)$$

D'altra parte si dimostra facilmente che è (2):

$$d = 2 \cdot {}^1S_M$$

---

(1) Questa uguaglianza vale anche se i due gruppi di $\frac{n}{2}$ termini, in cui si sono distinti gli n termini considerati, non sono composti l'uno dei termini superiori e l'altro dei termini inferiori alla mediana. Essa era stata già dimostrata in *Indici di omofilia e di rassomiglianza e loro relazioni col coefficiente di correlazione e con gli indici di attrazione* in «Atti del R. Istituto Veneto» ad. ord. del 25 ottobre 1914, Tomo LXXIV, Parte II, 1915, pagg. 585-586. In detta dimostrazione ci si riferiva però inesattamente alla differenza media, anzi che alla differenza media con ripetizione.

(2) Indichiamo con a_i $(i = 1, 2, \ldots\ldots, \frac{n}{2})$ un termine generico superiore e con a_k $(k = 1, 2, \ldots\ldots, \frac{n}{2})$ un termine generico inferiore alla mediana, sarà:

$$d = \frac{\sum_{i=1}^{\frac{n}{2}} \sum_{k=1}^{\frac{n}{2}} | a_i - a_k | + \sum_{k=1}^{\frac{n}{2}} \sum_{i=1}^{\frac{n}{2}} | a_k - a_i |}{\frac{n^2}{2}}$$

ossia, per essere i due addendi del numeratore del 2° membro manifestamente uguali tra loro,

$$d = 2 \frac{\sum_{i=1}^{\frac{n}{2}} \sum_{k=1}^{\frac{n}{2}} | a_i - a_k |}{\frac{n^2}{2}}$$

Ora, essendo ciascuno degli a_i superiore a tutti gli a_k, sarà, per qualunque a_i,

$$\sum_{k=1}^{\frac{n}{2}} | a_i - a_k | = \frac{n}{2} a_i - \sum_{k=1}^{\frac{n}{2}} a_k$$

e quindi, facendo assumere ad i tutti i valori da 1 a $\frac{n}{2}$,

così che si ricava:

$$\Delta_R = {}^1S_M + \frac{1}{4}(\Delta'_R + \Delta''_R)$$

vale a dire *la differenza media con ripetizione tra n termini è uguale allo scostamento semplice medio dalla mediana più un quarto della somma della differenza media tra gli* $\frac{n}{2}$ *termini superiori alla mediana e della differenza media tra gli* $\frac{n}{2}$ *termini inferiori alla mediana.*

Questa proposizione vale esattamente per un numero di termini pari ed approssimativamente per un numero di termini dispari.

$$\sum_{i=1}^{\frac{n}{2}} \sum_{k=1}^{\frac{n}{2}} |a_i - a_k| = \frac{n}{2} \sum_{i=1}^{\frac{n}{2}} a_i - \frac{n}{2} \sum_{k=1}^{\frac{n}{2}} a_k$$

Potremo allora scrivere:

$$d = 2\,\frac{\sum_{i=1}^{\frac{n}{2}} a_i - \sum_{k=1}^{\frac{n}{2}} a_k}{n} \qquad (1)$$

D'altra parte lo scostamento semplice medio dalla mediana:

$${}^1S_m = \frac{\sum_{i=1}^{\frac{n}{2}} (a_i - M) + \sum_{k=1}^{\frac{n}{2}} (M - a_k)}{n}$$

si può porre sotto la forma:

$${}^1S_m = \frac{\sum_{i=1}^{\frac{n}{2}} a_i - \frac{n}{2} M + \frac{n}{2} M - \sum_{k=1}^{\frac{n}{2}} a_k}{n}$$

o anche:

$${}^1S_m = \frac{\sum_{i=1}^{\frac{n}{2}} a_i - \sum_{k=1}^{\frac{n}{2}} a_k}{n}$$

Da quest'espressione e dalla (1) si ricava:

$$d = 2 \cdot {}^1S_m$$

CORRADO GINI

Del passaggio dall'indice di variabilità di un campione all'indice di variabilità della massa [1]

Ricorre spesso il problema di desumere, dall'indice di variabilità [2] di una collezione limitata di osservazioni, l'indice di variabilità di una collezione più ampia da cui la prima si ammette sia stata ricavata a caso.

Da molti la prima collezione si chiama « campione », e la denominazione mi pare appropriata ; la seconda « universo », e la denominazione mi pare meno felice. Poichè, in realtà, molto spesso una collezione di osservazioni può interessare per sè oppure come campione di una collezione più vasta, cosicchè bisognerebbe ammettere universi di diverso ordine [3], ciò che è contradditorio con lo stesso concetto di universo.

---

[1] Nota presentata alla riunione del 25 giugno 1946 del Seminario di Statistica dell'Istituto di Statistica della Università di Roma.

[2] In questa nota, quando parlo di « indice di variabilità » intendo « indice di variabilità assoluta ».

[3] I coscritti di un paese interessano l'autorità militare per i caratteri — qualunque ne sia l'origine — che essi effettivamente presentano ; essi costituirebbero, cioè, per essa, un « universo », mentre all'antropologo interessano come un « campione » da cui risalire ai caratteri della razza, quali risulterebbero in un numero infinitamente grande di individui cresciuti nelle stesse condizioni, ma sottratti all'azione delle circostanze accidentali che si fa sentire in un numero di osservazioni limitato.

Se, dei coscritti, prendiamo a considerare quelli che hanno estratto un numero alto di leva e in passato erano soggetti a una durata di ferma ridotta, otterremo un campione dell'universo dei coscritti e se, da quelli che hanno estratto un numero alto, scegliamo i coscritti che sono stati

In contrapposto al termine « campione », ritengo quindi preferibile usare un termine diverso da « universo ». Molte volte può essere appropriato usare il termine « partita » ; in generale, può usarsi — ed io qui userò — il termine generico di « massa ».(1)

* * *

Il problema del passaggio dall'indice di variabilità della massa al valore probabile dell'indice di variabilità del campione è risolto per varî indici di variabilità (scostamento quadratico medio, differenza media, differenza media con ripetizione) e comunemente si ritiene di avere così risolto anche il problema inverso del passaggio dall'indice di variabilità del campione al valore probabile dell'indice di variabilità della massa.

Essendo, ad esempio, dimostrato che è

$$E\,(^2S_p) = \frac{n}{n-1}\,\frac{m-1}{m}\,{}^2S_g$$

(dove E deve leggersi « valore probabile di » e dove n indica il numero dei casi della massa, m quelli del campione, ^2S_g lo scostamento quadratico medio della massa e ^2S_p lo scostamento quadratico medio del campione) e, per n molto grande,

$$E\,(^2S_p) = \frac{m-1}{m}\,{}^2S_g$$

ci si ritiene autorizzati ad usare la relazione

$$E\,(^2S_g) = \frac{n-1}{n}\,\frac{m}{m-1}\,{}^2S_p$$

---

scelti a caso per una data unità militare, otterremo un ulteriore campione, rispetto al quale i coscritti che hanno estratto un numero alto rappresentano un universo. Al comandante che esegue una statistica dei suoi soldati interessano i caratteri che i coscritti destinati alla sua unità effettivamente presentano, indipendentemente dal fatto che questi sieno più o meno dissimili da quelli dei soldati destinati ad altre unità. Perciò detto gruppo di coscritti verrebbe considerato da lui come un universo.

Totale dei coscritti, coscritti che hanno estratto un numero alto, coscritti destinati ad una data unità militare possono così, a seconda dei punti di vista, venire tutti considerati come universi o come campioni.

(1) Il termine « universo » può opportunemente riserbarsi ad una massa infinitamente grande. Cfr., nel volume XV di questa stessa Rivista, l'articolo *Le medie dei campioni*.

e, per n molto grande,

$$E\,(^2S_g) = \frac{m}{m-1}\,{}^2S_p$$

per risalire dallo scostamento quadratico medio del campione a quello della massa.

Non vi è, credo, statistico — lo scrivente non escluso (1) — che non abbia fatto uso di tali relazioni.

Il passaggio però non è legittimo senza il ricorso ad ipotesi ed io mi propongo appunto di precisare qui quali sono le ipotesi che lo legittimerebbero.

Il problema è analogo a quello che si pone per il passaggio dal valore di una grandezza assoluta osservata al suo valore vero, problema che ho trattato in apposita comunicazione presentata, nel giugno 1943, alla Società Italiana di Statistica (2) e può anzi dirsi che il primo problema non sia se non un caso particolare del secondo.

* * *

Condizioni sufficienti per che sia legittimo il passaggio dall'indice di variabilità del campione all'indice di variabilità della massa è, in base ai risultati di detta memoria:

a) che tutti i valori dell'indice di variabilità della massa si possano riguardare a priori — prima, cioè, dell'osservazione del fenomeno — come ugualmente possibili;

b) che, qualunque sia il valore dell'indice di variabilità, la distribuzione degli errori sia identica.

---

(1) Vedasi, in particolare, la memoria *Variabilità e mutabilità*, Cagliari, 1912, riprodotta in *Memorie di metodologia statistica*, Giuffrè, 1935, in particolare al capitolo: *Degli indici di variabilità dei caratteri nel caso di seriazioni parziali*, pagg. 228-236.

(2) *Sulla probabilità inversa nel caso di grandezze a distribuzione costante*, in «Atti della VI e VII Riunione della Società It. di Statistica», 1945, pagg. 283-299.

Questa seconda ipotesi è evidentemente inammissibile.

Se si può ammettere, invero, che, al variare del valore dello indice di variabilità, la *forma* della distribuzione degli errori resti costante, in quanto detta forma dipende essenzialmente dalla natura del fenomeno, non si può ammettere, invece, che resti costante l'*ampiezza* della sua distribuzione, chè evidentemente — se l'indice di variabilità corrisponde allo scopo — essa deve crescere o diminuire, a parità di forma, proporzionalmente al crescere o diminuire dell'indice di variabilità.

Ma, se l'ampiezza della distribuzione cresce o diminuisce proporzionalmente all'indice di variabilità, resterà costante la distribuzione, non degli errori, ma dei logaritmi degli errori. Ora, se noi ammettiamo tale ipotesi ed ammettiamo contemporaneamente che, a priori, sieno equiprobabili, non tutti i valori dell'indice di variabilità, ma tutti i valori del logaritmo di tale indice, il passaggio in questione diventa legittimo.

Riportati infatti sull'asse delle ascisse della figura a pag. 284 della comunicazione sopra citata, non i valori della grandezza, ma i rispettivi logaritmi, la dimostrazione corre ugualmente, ma implica l'ipotesi che sieno equiprobabili a priori i logaritmi di tutti i valori della grandezza vera che si vuole determinare e che, al variare di tali valori rimanga costante la distribuzione dei logaritmi dei valori osservati.

In detta memoria, io avvertivo a pag. 290, che la costanza della distribuzione dei valori osservati forse non si verifica mai e in particolare osservavo che non è ammissibile che la distribuzione delle stature osservate sia indipendente dalla statura tipica. A molti sembra invece ragionevole ammettere che sia indipendente dalla statura tipica la distribuzione dei logaritmi delle stature; ed in fondo questa è l'ipotesi implicita nell'uso del coefficiente di variazione (rapporto dell'indice di variabilità alla media aritmetica) al fine di paragonare la variabilità delle stature, come quella di tanti altri caratteri, di popolazioni diverse.

Senza entrare nel merito del fondamento di tale ipotesi, basti qui fermare il risultato generale che il passaggio, dal valore osservato di una grandezza osservata nel campione al valore ignoto della grandezza vera nella massa, da cui il campione è ricavato — passaggio ottenuto invertendo la relazione che passa tra il valore ignoto della grandezza vera nella massa e il valore noto

della grandezza nel campione — risulta legittimo, non solo quando si ammettono le condizioni seguenti:

a) equidistribuzione delle probabilità a priori dei vari valori della grandezza nella massa;

b) invarianza della distribuzione dei valori della grandezza nel campione al variare del valore della grandezza nella massa; come è stato da me dimostrato nel 1943,

ma anche quando si ammettono le due condizioni seguenti:

α) equidistribuzione delle probabilità a priori dei logaritmi dei vari valori della grandezza nella massa;

β) invarianza della distribuzione dei logaritmi dei valori della grandezza nel campione al variare del valore della grandezza nella massa.

G. DAHLBERG

Eine Methode zur Messung von Rassenunterschieden.

Innerhalb der Anthropologie hat man bisher in verhältnismässig begrenztem Umfang statistische Methoden angewandt. Wenn man Rassenunterschiede untersucht hat, hat man sich im grossen und ganzen mit Mittelwert (arithmetischem Mittel), mittlerem Fehler sowie Variabilitätsmass begnügt, ohne darüber hinaus eine nähere Analyse anzustreben. Man hat zu unterscheiden versucht, ob die Differenzen zwischen den Mittelwerten statistisch sichergestellt waren oder nicht. Man hat sich deshalb im allgemeinen mit der Feststellung begnügt, dass zwischen bestimmten Gruppen Unterschiede vorliegen, die unbestreitbar sind, aber man besitzt keine Methode, die einen Vergleich der Unterschiede zwischen verschiedenen Gruppen ermöglicht. Man ist sich natürlich darüber im Klaren, dass z. B. der Unterschied zwischen Neger und Mongole — um ein extremes Beispiel zu gebrauchen — viel grösser ist als der Unterschied zwischen europäischen Rassen, aber man kann nicht angeben, um wieviel grösser der zuerstgenannte Unterschied ist. Der Pearson'sche « coefficient of racial likeness » ist bei der Entscheidung zu verwenden, ob ein Unterschied vorliegt oder nicht. Der Koeffizient hat folgende Form :

$$C.R.L. = \frac{1}{m}\left\{\Sigma\ \frac{(M_i' - M_i'')^2}{\varepsilon'^2 + \varepsilon''^2}\right\} - 1 \pm \frac{0.67449}{\sqrt{2\,m}}$$

wobei m die Anzahl Eigenschaften angibt, womit der Vergleich zwischen den Rassen vorgenommen wird. $M' \pm \varepsilon'$ und $M'' \pm \varepsilon'$, drückt die Mittelwerte und mittleren Fehler der Eigenschaften aus.

Man sieht ohne weiteres, dass der Koeffizient grösser wird, wenn die Anzahl untersuchter Personen wächst. Der Koeffizient ist ja so aufgebaut, dass man mit dem mittleren Fehler, der ja

der Wurzel aus der Anzahl untersuchter Individuen proportional ist, wägt.

Um zu einem befriedigenden Mass für Rassenunterschiede zu gelangen, kann man von folgendem Gedankengang ausgehen. Man analysiert zuerst die Eigenschaften, von denen man bei der Analyse ausgehen will und berechnet den Mittelwert und die

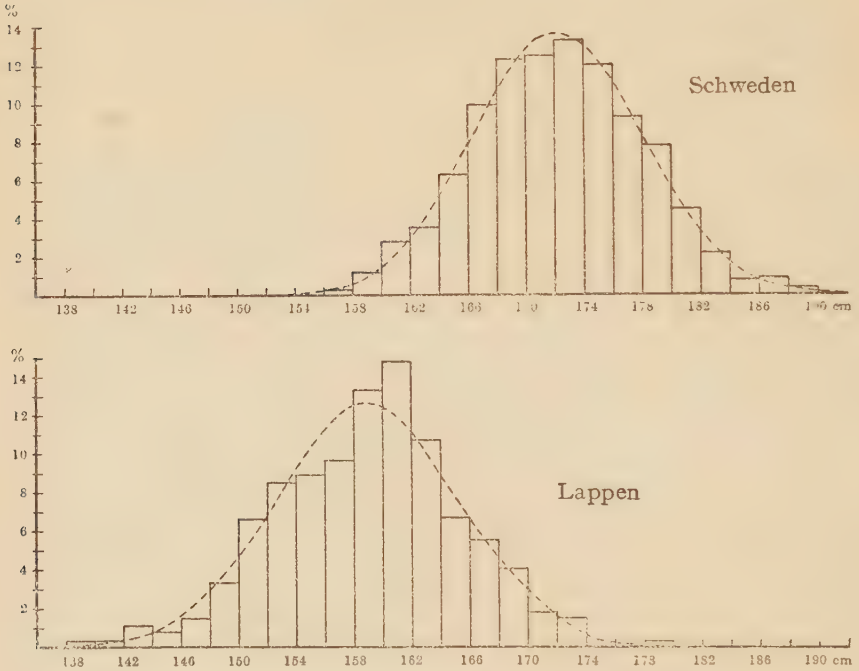


Diagramm 1. Verteilungskurve für 1000 Schweden und 269 Lappen bezüglich ihrer Körperhöhe (Männer).

Variationsbreite. Darnach macht man wieder das Gesamtmaterial zur Grundlage der weiteren Untersuchung wobei man mit Hilfe der erworbenen Kenntnisse eine Aufbereitung nach den getrennten Gruppen vornimmt. Man kann dabei zwei Wege gehen. Man sortiert zuerst nach einzelnen Eigenschaften, darnach nach Eigenschaftskombinationen. Die Zuteilung wird jeweils durch die vorher bestimmten Variabilitätsgrenzen entschieden. Auf diese Weise kann man feststellen, einen wie grossen Teil der entsprechenden Gruppen man mit Sicherheit der einen oder anderen Gruppe zuteilen kann. Die Grösse dieses An-

teils ist ein Mass für die Grösse des Rassenunterschiedes. Dieser Unterschied muss offensichtlich gross sein, wenn man eine sichere Rassendiagnose für eine grosse Anzahl Individuen beider Materialien stellen kann, und umgekehrt. Man kann ferner derart vorgehen, dass man diejenige Sortierung vornimmt, die wahrscheinlich das beste Ergebnis zeitigt, und dann bestimmt, in welchem Ausmass man dabei eine fehlerhafte Sortierung vorgenommen hat. Wenn die fehlerhaften Diagnosen verhältnismässig häufig vorkommen, sind die Rassenunterschiede klein und umgekehrt.

Mit diesen Gesichtspunkten als Ausgangspunkt hat der Verfasser eine Methode ausgearbeitet, die für eine Analyse des Rassenunterschiedes zwischen Schweden und Lappen verwendet wird (vgl. G. Dahlberg: *The conception of race and a new method of delimiting races, demonstrated on a material of Swedes and Lapps* in: G. Dahlberg und S. Wahlund: *The race biology of the Swedish Lapps.* Part II. (Uppsala 1941).

Will man zwei Populationen (P' und P'') bezüglich einer quantitativ messbaren Eigenschaft A (A' und A'') vergleichen, die «normal» um ihren Mittelwert variieren, so kann man die Abweichungen jedes Individuums von dem Mittelwert der Mittelwerte der beiden Populationen (M'_A und M''_A) bestimmen, ausgedrückt in der Dispersion der Bevölkerung, der das Individuum angehört. Wir können die Individuen der beiden Serien auf folgende Weise ausdrücken:

1) Serie P':

$$\frac{A'_i - \frac{M'_A + M''_A}{2}}{\sigma'_A}$$

Der Mittelwert der Serie =

$$= \mathfrak{M}'_A = \frac{1}{n'} \sum_{i=1}^{n'} \frac{A'_i - \frac{M'_A + M''_A}{2}}{\sigma'_A} =$$

$$= \frac{M'_A - M''_A}{2\,\sigma'_A}$$

Die Dispersions quadrate $\sigma_A'^2 =$

$$= \frac{1}{n'} \sum_{i=1}^{n'} \left[\frac{A_i' - \frac{M_A' + M_A''}{2}}{\sigma_A'} - \frac{M_A' - M_A''}{2\,\sigma_A'} \right]^2 =$$

$$= \frac{1}{n'} \sum_{i=1}^{n'} \left[\frac{A_i' - M_A'}{\sigma_A'} \right]^2 = 1 .$$

2) Serie P'':

$$\frac{A_i'' - \frac{M_A' + M_A''}{2}}{\sigma_A''}$$

Der Mittelwert der Serie $=$

$$= \mathfrak{M}_A'' = \frac{1}{n''} \sum_{i=1}^{n''} \frac{A_i'' - \frac{M_A' + M_A''}{2}}{\sigma_A''} =$$

$$= \frac{M_A'' - M_A'}{2 \cdot \sigma_A''}$$

und die Dispersion $\sigma_A'' = 1$ (siehe oben).

Insofern die beiden Populationen verschiedene Mittelwerte haben, erhält man eine positive oder negative Zahl als Ausdruck für die neuen Mittelwerte. Oft hat die untersuchte Eigenschaft bei beiden Populationen die gleiche Dispersion, weswegen $\mathfrak{M}'$ und $\mathfrak{M}''$ numerisch der gleichen Grössenordnung angehören.

Werden N Eigenschaften kombiniert, so erhält man für die beiden Populationen die Mittelwerte $\mathfrak{M}_s'$ und $\mathfrak{M}_s''$.

$$\mathfrak{M}_s' = \sum_{i=1}^{N} \frac{\left| M_i' - M_i'' \right|}{2 \cdot \sigma_i'} \qquad (1\,a)$$

$$\mathfrak{M}_s'' = \sum_{i=1}^{N} \frac{\left| M_i'' - M_i' \right|}{2 \cdot \sigma_i''} \qquad (1\,b)$$

Da die Dispersion σ_A für jede Eigenschaft 1 ist, so muss, falls man bei jedem Individuum N Eigenschaften kombiniert, unter der Voraussetzung, dass zwischen den Eigenschaften keine Korrelation vorliegt, die Dispersion in einer solchen Serie $\sigma_s = \sqrt{N}$ werden.

Besteht dagegen zwischen den Eigenschaften ein Zusammenhang, so muss dies natürlich eine Erhöhung des Dispersionswertes mitsichführen. Bezeichnet man die durchschnittliche Korrelation

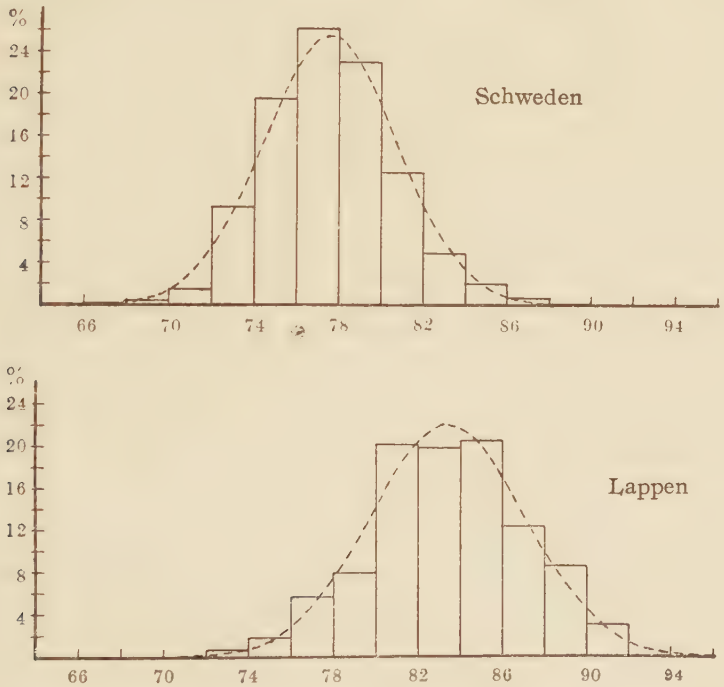


Diagramm 2. Verteilungskurve fur 1000 Schweden und 269 Lappen bezüglich ihres Kopfindexes.

zwischen den beobachteten Eigenschaften in den Populationen P' und P'' mit r' resp. r'', so erhält man die Dispersion wie folgt:

$$\sigma_s'^2 = N + N \,.\, (N - 1)\, r' \text{ resp.} \qquad (2\,a)$$

$$\sigma_s''^2 = N + N \,.\, (N - 1)\, r'' \qquad (2b)$$

Hierbei muss bemerkt werden, dass wir bei der Kombination mehrerer Eigenschaften angenommen haben, dass die Ausdrücke aus der die Summen $\mathfrak{M}_s'$ und $\mathfrak{M}_s''$ gebildet sind, gleiche Vorzeichen zeigen. Mann nimmt also kein Rücksicht darauf, dass die Mittelwerte der Eigenschaften in der einen Population bald grösser, bald kleiner als in der anderen sind. Man kann die Sache auch so ausdrücken, dass man die Ausdrücke ohne Rücksicht auf das Vorzeichen summiert und den ersten Ausdrück das Vorzeichen der Summe bestimmen lässt.

Wenn die beobachteten Eigenschaften korreliert sind, sieht man, dass eine grosse Abweichung von $A : s$ Mittelwert $\left(\text{oder genauer } \frac{M_A' + M_A''}{2}\right)$ auch im allgemeinen eine grosse Abweichung von $B : s$ Mittelwert $\frac{M_B' + M_B''}{2}$ nach sich zieht. Addiert man die Eigenschaften A und B, so wird sich dadurch die Dispersion der Kombination vergrössern. Dies gilt natürlich sowohl bei positiver wie bei negativer Korrelation. Bei der Berechnung der Dispersion lässt man somit das Vorzeichen der Korrelationskoeffizienten unberücksichtigt.

Folgendes Beispiel beleuchtet näher, wie man bei den Berechnungen zu Wege geht.

Bei einer anthropometrischen Untersuchung von 269 reinen Lappen (Männer) und 1000 schwedischen Rekruten erhielt man für die Körperhöhe, den Kopfindex und den morphologischen Gesichtsindex die in Tabelle 1 wiedergegebenen Werte. Die Verteilungskurven sind normal (siehe Diagramme 1-3).

TABELLE 1. — *Mittelwert und Dispersion der Körperhöhe, des Kopf- und morphologischen Gesichtsindexes für* 269 *Lappen und* 1000 *Schweden.*

Eigenschaft	Lappen		Schweden	
	$M \pm \varepsilon(M)$	σ	$M \pm \varepsilon(M)$	σ
Körperhöhe	158.9±0.4	6.4	172.2±0.2	5.8
Kopfindex.	83.4±0.2	3.6	77.6±0.1	3.1
Morph. Gesichtsindex.	83.0±0.3	5.3	93.0±0.2	5.7

In der Lappenpopulation sind die Werte für die Körperhöhe und den Gesichtsindex kleiner als für die « Schweden » population. Dagegen ist der Kopfindex der Lappen höher als bei den Schweden. Der Mittelwert der Lappen $\left(\text{gemäss } \mathfrak{M}' = \frac{M' - M''}{2\,\sigma}\right)$ wird also bezüglich Körperhöhe und Gesichtsindex negativ, bezüglich des Kopfindex positiv. Für die Schweden gilt das entgegengesetzte (siehe Tabelle 2).

TABELLE 2. — *Mittelwert und Dispersion gemäss der Formeln* 1 *und* 2.

Eigenscnaft	Lappen		Schweden	
	$M \pm \varepsilon(M)$	σ	$M \pm \varepsilon(M)$	σ
Körperhöhe	—1.05±0.06	1.00	+1.15±0.03	1.00
Kopfindex.	+0.79±0.06	1.00	—0.94±0.03	1.00
Morph. Gesichtsindex.	—0.94±0.06	1.00	+0.87±0.03	1.00
Körperhöhe und Kopfindex. .	—1.84±0.09	1.50	+2.09±0.05	1.46
Körperhöhe, Kopf- u. morph.. Gesichtsindex	—2.78±0.12	2.00	+2.96±0.06	1.93

Bei der Addition der Mittelwerte von zwei oder drei Eigenschaften (worin der Hauptindex enthalten ist) wird das Vorzeichen vernachlässigt.

Wir erhalten also bei der Kombination :

1) Körperhöhe und Kopfindex bei Lappen

$$\mathfrak{M}'_s = -(1.05 + 0.79) = -1.84$$

2) Körperhöhe und Kopfindex bei Schweden

$$\mathfrak{M}''_s = +(1.15 + 0.94) = +2.09\,.$$

Auf die gleiche Art und Weise erhält man für alle drei Eigenschaften :

$$\mathfrak{M}'_s = -2.78$$

$$\mathfrak{M}''_s = +2.96\,.$$

Zwischen Körperhöhe und Kopfindex sowie zwischen Kopfindex und Gesichtsindex liegt negative, zwischen Körperhöhe und Gesichtsindex dagegen positive Korrelation vor. Die Werte dieser Korrelationskoeffizienten finden sich in Tabelle 3.

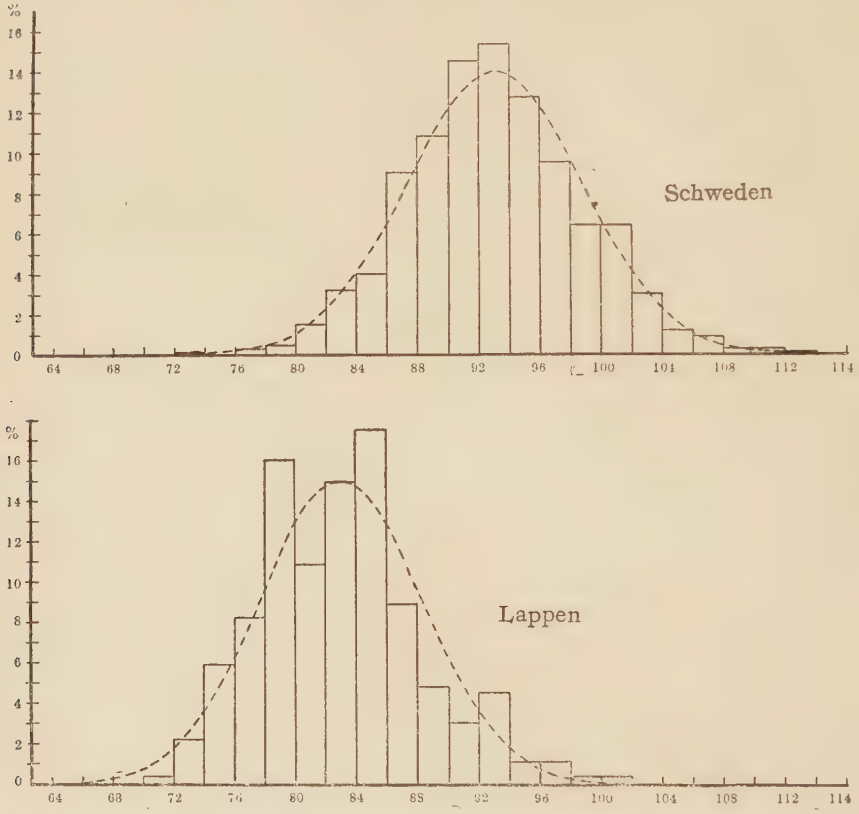


Diagramm 3. Verteilungskurve für 1000 Schweden und 269 Lappen bezüglich ihres morphologischen Gesichtsindexes.

Tabelle 3. — *Korrelationskoeffizient zwischen Körperhöhe, Kopfindex und morphologischem Gesichtsindex für Lappen und Schweden.*

Korrelation zwischen	Lappen	Schweden
Köperhöhe und Kopfindex	—0.16±0.06	—0.07±0.03
Körperhöhe und morph. Gesichtsindex.	+0.09±0.06	+0.13±0.03
Kopfindex und morph. Gesichtsindex	—0.27±0.06	—0.16±0.03

Die durchschnittliche Korrelation beträgt in vorliegenden Material für die Lappen $(0.16 + 0.09 + 0.27)/3 = 0.17$ und für die Schweden $(0.07 + 0.13 + 0.16)/3 = 0.12$. Mittels dieser Zahlen können wir die Dispersion bei der Kombination der 3 Masse berechnen: $\sigma = \sqrt{3 + 6 \cdot 0.17} = 2.00$ für die Lappen und $\sigma = 1.93$ für die Schweden.

Die in Tab. 2 angegebenen Resultate sind durch Berechnung des « Koeffizienten » für jedes einzelne Individuum beider Populationen bestätigt worden. Die Mittelwerte und Dispersionen der beiden Verteilungskurven sind berechnet worden und stimmten gut mit den in Tab. 1 angegebenen Werten überein. Wir teilen die Werte der erhaltenen Verteilungskurven hier nicht mit, da sie kaum von Interesse sind.

Die schwedische und lappische Verteilungskurve kombinierter Werte zeigen beide die Form der Normalkurve und verschieben sich, je mehr die Anzahl beobachteter Eigenschaften zunimmt, längs der Abszisse in entgegengesetzter Richtung. Je grösser die Anzahl beobachteter Eigenschaften wird, desto stärker wird die Verschiebung. Bei *einer* Eigenschaft, z. B. der Körperhöhe, beträgt der Abstand zwischen den Mittelwerten 2.20 (= 1.05 + + 1.15, siehe Tab. 2), was 37% der Variationsbreite 6σ (= 6) gleichkommt. Bei der Kombination Körperhöhe und Kopfindex ist der gerade genannte Abstand bis auf 3.93 gewachsen, was 44% der Variationsbreite 9 (= 6×1.5) gleichkommt. Bei der Kombination aller drei Masse beträgt der Abstand zwischen den Mittelpunkten der Verteilungskurven 48% der Variationsbreite, die 12 beträgt. Bei genügend grosser Anzahl Eigenschaften kann man sich also denken, dass die Verschiebung bis zum Betrag der ganzen Variationsbreite wächst, dies jedoch unter bestimmten Bedingungen bezüglich der durchschnittlichen Zunahme des Mittelwertes für jede Eigenschaft, sowie der Grösse der Korrelation. Im Grenzfall liegen die beiden Verteilungskurven neben einander, sodass der niedrigste Wert ($M' - 3\sigma'$) der einen Verteilungskurve mit dem höchsten Wert der anderen ($M'' + 3\sigma''$) zusammenfällt. Vermehrt man die Anzahl Eigenschaften noch mehr, so trennen sich die beiden Kurven. Der gerade genannte Grenzfall schliesst deutlich in sich, dass man hier die Möglichkeit hat mit Bestimmtheit eine anthropologische *Individualdiagnose* zu stellen. Ein Individuum, *von dem man im Voraus weiss, dass es zu einer* (*aber nicht welcher*) *der beiden Populationen gehört*, kann

man nach Untersuchung einer bestimmten Anzahl Eigenschaften mit Gewissheit der einen oder anderen Population zuteilen. Man erreicht nicht mehr, wenn man weitere Eigenschaften beobachtet.

Wenn man, um die Sache zu vereinfachen, annimmt, dass die beiden zu vergleichenden Populationen für eine gewisse Eigenschaft die gleiche Dispersion haben, so werden die Mittelwerte natürlich die gleichen numerischen Werte haben (der eine negativ,

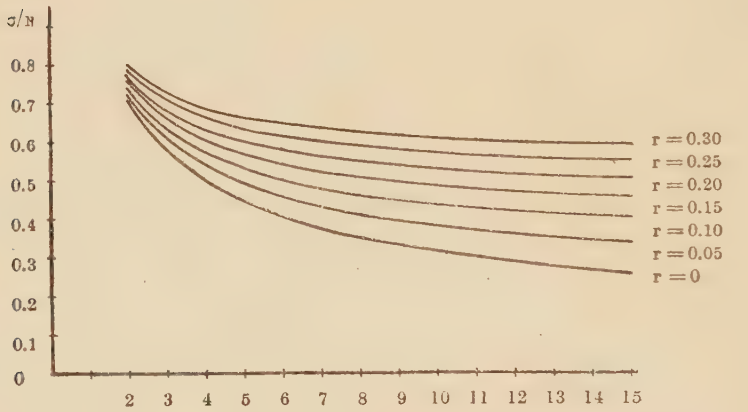


DIAGRAMM 4. Quotient $\frac{\sigma}{N}$ bei verschieden grosser Korrelation und verschiedener Anzahl beobachteter Eigenschaften.

der andere positiv). Nimmt man ferner an, dass die Korrelation innerhalb der beiden Gruppen die gleiche ist, so werden natürlich auch die Dispersionen um diese Mittelwerte gleich. Diese Dispersion ist am kleinsten, wenn keine Korrelation zwischen den beobachteten Eigenschaften vorliegt und nimmt bei steigender Korrelation zu, unabhängig, ob diese positiv oder negativ ist. Die Grösse der Dispersion bei verschiedener durchschnittlicher Korrelation und verschiedener Anzahl beobachteter Eigenschaften geht aus Tab. 4 und dem Diagramm 4 hervor.

Vergleicht man Tab. 2 mit Diagramm 5, so ergibt sich u. a. folgendes:

Der Abstand zwischen den Mittelwerten der beiden Verteilungskurven (Populationen) beträgt $2.96 + 2.78 = 5.74$, was genau 3 mal der Dispersion entspricht, wenn wir diese mit 1.9

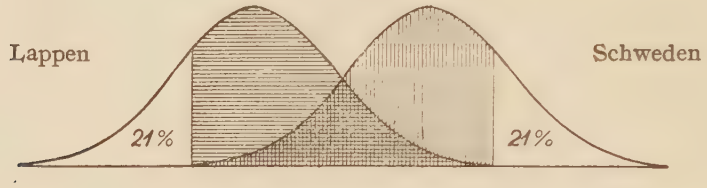
annehmen (approximativ richtig, siehe Tab. 2). Dies schliesst in sich, dass jede der beiden Verteilungskurven im Mittelpunkt (Mittelwert) der anderen Kurve endet. Wir können mit anderen Worten sagen, dass in sehr seltenen Fällen ein Schwede den für Lappen typischen Wert (Mittelwert) aufweist und umgekehrt. Individuen mit niedrigerem Wert als — 2.78 können wir mit

TABELLE 4. — *Quotient* $\frac{\sigma}{N}$ *bei verschieden grosser Korrelation gemäss der Formel* $\frac{\sigma}{N} = \sqrt{\frac{N+(N-1).r}{N^2}}$, *worin* $N =$ *Anzahl Eigenschaften.*

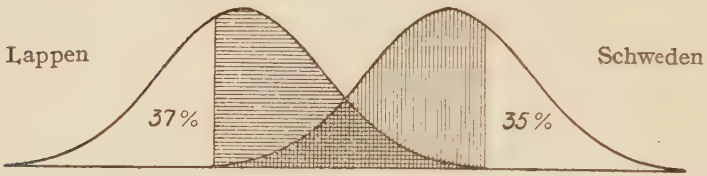
N	σ/N						
	$r = 0$	$r = 0.5$	$r = 0.10$	$r = 0.15$	$r = 0.20$	$r = 0.25$	$r = 0.30$
2	0.71	0.72	0.74	0.76	0.77	0.79	0.81
3	0.58	0.61	0.63	0.66	0.68	0.71	0.73
4	0.50	0.54	0.57	0.60	0.63	0.66	0.69
5	0.45	0.49	0.53	0.57	0.60	0.63	0.66
6	0.41	0.46	0.50	0.54	0.58	0.61	0.65
7	0.37	0.43	0.48	0.52	0.56	0.60	0.63
8	0.36	0.41	0.46	0.51	0.55	0.59	0.62
9	0.33	0.39	0.45	0.49	0.54	0.58	0.61
10	0.32	0.38	0.44	0.48	0.53	0.57	0.61
11	0.30	0.37	0.43	0.48	0.52	0.56	0.60
12	0.29	0.36	0.42	0.47	0.52	0.56	0.60
13	0.28	0.35	0.41	0.46	0.51	0.55	0.59
14	0.27	0.34	0.41	0.46	0.51	0.55	0.59
15	0.26	0.34	0.40	0.45	0.50	0.55	0.59
∞	0	0.22	0.32	0.39	0.45	0.50	0.55

Gewissheit als Lappen bezeichnen, ebenso müssen Individuen mit höherem Wert als + 2.96 Schweden sein. Individuen mit Werten zwischen den gerade genannten Werten können wir nicht mit *absoluter* Sicherheit diagnostizieren. Es ist doch klar, dass je niedriger der Wert eines Individuums innerhalb des Intervalls + 2.96 und — 2.78 wird, desto grösser wird die Wahrscheinlichkeit, dass es Lappe ist. Wenn wir vom Mittelwert der Schweden + 2.96 ausgehen und gegen niedrigere Werte gehen, können wir sagen, dass zu Anfang die Wahrscheinlichkeit für Schwedisch grösser ist als die für Lappisch. Hier liegen ja nur die allerhöchsten Werte der Verteilungskurve der Lappen. Wir kommen jedoch zu einem Wert, wo die

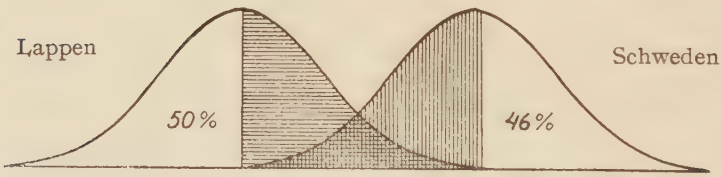
Wahrscheinlichkeit für Schwedisch genau so gross wird wie für Lappisch (die Ordinaten der beiden Kurven sind gleich). Dieser Wert wird im Diagramm durch den Schnittpunkt der beiden Verteilungskurven (Normalkurven) representiert und beträgt un-



Körperhöhe,



Körperhöhe und Kopfindex kombiniert



Körperhöhev, Kopfinder u. morph. Gesichtsindex kombiniert

DIAGRAMM 5. Verteilungskurbe der Masse gemäss Formel 1 für Lappen und Schweden.

gefähr + 0.10. Unterhalb dieses Wertes liegen theoretisch (gemäss der Wahrscheinlichkeitsfunktion) 6.7% der Schweden. Ein gleich grosser Prozentsatz Lappen liegt oberhalb des genannten Wertes. Geht man von den individuellen Werten der beiden Populationen aus, worauf wir ja unsere Berechnungen aufbauten, so finden wir unterhalb des Grenzwertes 7.9% Schweden und oberhalb des-

selben 8.2% Lappen. Die gemeinsame Frequenzfläche betrug demnach 16.1 (± 1.9) %, entgegen der theoretisch berechneten 13.4 (= 2 × 6.7%). Die Abweichung liegt innerhalb der mittleren Fehlergrenze. Unterhalb dieses Punktes auf der Abszisse (+ 0.10) wendet sich die Wahrscheinlichkeit und wird grösser für Lappisch und kleiner für Schwedisch. Wenn wir zum Wert — 0.8 kommen, befinden wir uns 2σ vom schwedischen Mittelwert + 2.96 entfernt. Unterhalb — 0.8 liegen also theoretisch 2.3% der Schweden (beobachteter Wert 2.3%). Wenn wir annehmen, dass die beiden Verteilungen die gleiche Streuung haben, sollte dieser Punkt $M + 1\sigma$ der Verteilungskurve für die Lappen entsprechen. Oberhalb dieses Wertes liegen theoretisch 15.9% Lappen (der beobachtete Wert beträgt 14.9%). Beim Grenzwert — 0.8 wird also der totale Fehler grösser als beim Grenzwert + 0.10 (Kurvenschnittpunkt). Es verhält sich tatsächlich so, dass der Fehler sein Minimum dann erreicht, wenn die Grenze zwischen den beiden Verteilungen auf den Punkt der Abszisse verlegt wird, der dem Kurvenschnittpunkt entspricht. Haben wir es mit einerPopulation zu tun, die aus Schweden und Lappen besteht, und sollen wir die Individuen mittels der Masse für die Körperhöhe, den Kopfindex und den morphologischen Gesichtsindex den beiden Teilgruppen zuteilen, so erzielen wir das sicherste Resultat, wenn wir uns für den Wert + 0.10 als Grenzwert entscheiden. Alle Individuen mit niedrigerem Wert betrachten wir als Lappen, alle mit höherem Wert als Schweden. Dabei haben wir dann einen Fehler begangen, der sich auf 6-7% beläuft. In der Gruppe « Lappen » finden sich dann 6-7% Schweden und in der Gruppe « Schweden » der gleiche Prozentsatz der wirklichen Lappenanzahl. Dies schliesst jedoch *nicht* in sich, dass wir eine sichere Individualdiagnose für 93% jeder Population stellen können. Nein, die Diagnose ist für alle Individuen zwischen den Werten — 2.78 und + 2.96 mehr oder weniger unsicher (siehe S. 139). Am unsichersten ist sie natürlich für Individuen mit Werten um den Grenzwert. Diese Individuen können gleich gerne der einen wie der anderen Teilpopulation angehören.

Nun darf man nicht vergessen, dass, falls die beiden zu vergleichenden Gruppen keine Unterschiede aufweisen, bei einem Scheidungsversuch 50% richtig und 50% falsch sortiert werden. Davon ausgehend kann man für die Rassenscheidung folgen-

den Rassenindex R_2 aufstellen :

$$R_2 = \frac{p - 50}{50}$$

worin p die Anzahl Personen ausdrückt, die richtig sortiert wurden.

Unter bestimmten Annahmen bezüglich der durchschnittlichen Zunahme des Mittelwertes für jede neue beobachtete Eigenschaft, kann man die niedrigste Anzahl Eigenschaften berechnen, die beobachtet werden muss, auf dass sich die beiden Verteilungskurven so weit von einander verschieben, dass sie sich praktisch nicht berühren (die Variationsbreite wird $= 6\,\sigma$ gesetzt).

Wenn man die durchschnittliche Zunahme mit a bezeichnet, die Anzahl Eigenschaften mit N und die durchschnittliche Korrelation zwischen diesen Eigenschaften (unabhängig vom Vorzeichen der Korrelationskoeffizienten) mit r, so kann die Minimumanzahl Eigenschaften aus der folgenden Formel berechnet werden :

$$2\,a\,N = 6\,\sqrt{N + N\,(N - 1)\,r},$$

woraus man erhält

$$N = \frac{9\,(1 - r)}{a^2 - 9\,r}\ ;$$

Gemäss dieser Formel ist N für $a = 0.7,\ 0.8,\ 0.9,\ 1.0,\ 1.1,\ 1.2,\ 1.4\ ,\ 1.6$ und 1.8 berechnet worden (siehe Diagramm 6). Wir können schreiben :

$$a = \frac{1}{N} \sum_{i=1}^{N} \frac{M_i' - M_i''}{2\,\sigma_i}\ ,$$

woraus man erhält

$$M' - M'' \backsim 2\,a\,.\,\sigma$$

was in sich schliesst, dass, wenn a *z. B.* $= 1$, so muss die Differenz zwischen den Mittelwerten der beiden Populationen im allgemeinen doppelt so gross sein wie die Dispersion der zur Diskussion

stehenden Eigenschaft. In vorliegendem Beispiel liegt a bei ungefähr 1. Man kann vielleicht behaupten, dass die Differenz zwischen den Mittelwerten der beiden Gruppen relativ gross ist, wenn diese die doppelte Dispersion ausmacht, weshalb man

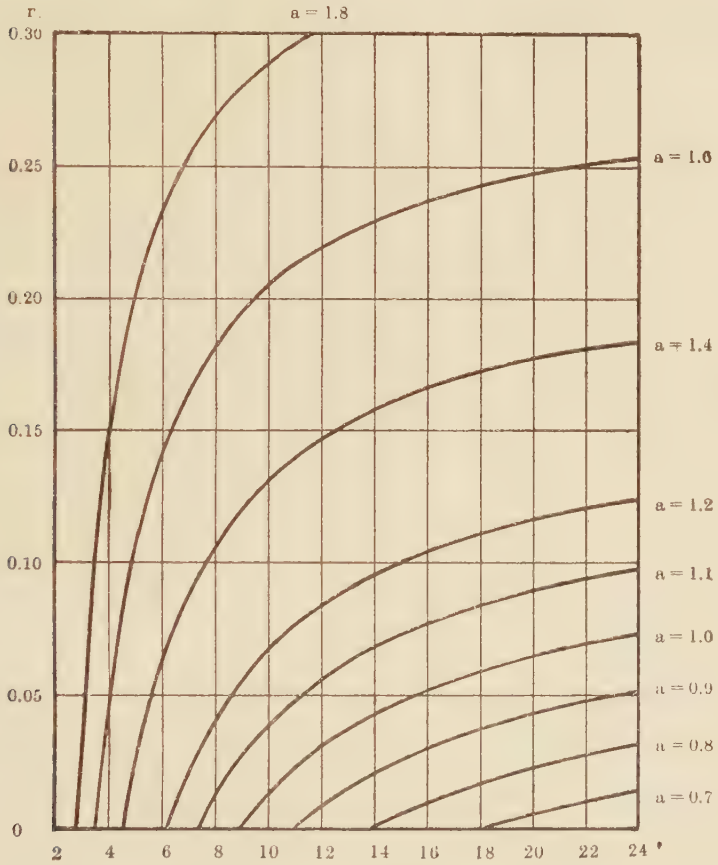


Diagramm 6. Die minimale Anzahl Eigenschaften, die herangezogen werden müssen, um jedes Individ bei verschieden grosser durchschnittli cher Korrelation (r) und durchschnittlichem Abstand der Mittelwerte (a) bestimmen zu können.

erwarten könnte, dass nicht allzuviele Eigenschaften beobachtet werden müssen, damit der Abstand zwischen den Mittelwerten der beiden zu vergleichenden Gruppen 6σ erreicht. In dem hier behandelten Fall beträgt die durchschnittliche Korrelation bei

den Lappen 0.17 und 0.12 bei den Schweden. Nun verhält es sich indessen so, dass die *durchschnittliche* Korrelation mit steigender Eigenschaftszahl fällt. Nehmen wir an, dass die Korrelation bis hinab zum Wert 0.05 sinkt, so geht aus dem Diagramm 6 hervor, dass bei $a = 1$ ($M' - M'' = 2\sigma$) wenigstens 16 Eigenschaften beobachtet werden müssen um völlige Gewissheit bei der Individualdiagnose zu erhalten. Bei O-Korrelation (welche in vorliegendem Fall unangemessen ist, nachdem eine durchschnittliche Korrelation von ca. 0.15 zwischen Körperhöhe, Kopf- und Gesichtsindex faktisch bewiesen ist) ist die geringste Anzahl Eigenschaften 9.

Aus dem Ausgeführten ersieht man, dass ein « coefficient of racial difference », um für die Bestimmung einer *Individualdiagnose* anwendbar zu sein, aufgebaut sein muss auf :

1) Eigenschaften, die inbezug auf ihre Dispersionen grosse Differenzen zwischen den zu vergleichenden Populationen aufweisen ;

2) Eigenschaften, die unter sich in einer möglichst kleinen Korrelation stehen.

Man muss also relativ viele Eigenschaften, mindestens 9 bei a = 1 untersuchen, auf dass sich die Verteilungskurven so weit von einander entfernen, dass der Abstand zwischen den Mittelwerten der ganzen Variationsbreite gleich wird. Da wir uns in dem hier behandelten Beispiel mit 3 Maassen begnügen, können wir erwarten, dass ein relativ grosser Teil der beiden Verteilungsflächen gemeinsam ist. Bei der Kombination von Körperhöhe und Kopfindex macht die gemeinsame Fläche 17% aus und bei der Kombination der gerade genannten sowie dem morphologischen Gesichtsindex 13%. Der Abstand zwischen den Mittelwerten beträgt im ersten Falle 44% und im letzteren 48% der Variationsbreite (siehe S. 141). Das Diagramm veranschaulicht das folgende. Unterhalb $M - 2\sigma$, gerechnet vom Schwedenmittelwert für die Kombination Körperhöhe und Kopfindex liegen 71% Lappen, unterhalb $M - 2\frac{1}{2}\sigma$ 58% sowie endlich unterhalb $M - 3\sigma$ 37% Lappen. Die entsprechenden Zahlen für die Kombination aller drei Masse sind 80, 71 und 50% Lappen. Wie erwartet, steigen die Prozentzahlen, aber 100% -ige Sicherheit erhält man bei der Diagnose nicht. Unterhalb eines Wertes des « Koeffizienten » von — 0.84 ($= M_{00} - 2\sigma$) liegen 80 % Lappen, aber auch 2.3% Schweden und das

Risiko für eine Fehldiagnose beträgt somit 3%. Praktisch berührt dieses Risiko das Intervall von — 0.8 bis — 2.8, weswegen das Risiko für eine Fehldiagnose innerhalb dieses Intervalls 7% beträgt. Kleinere Werte als — 2.8 finden sich nur unter Lappen, höhere Werte als 3.0 nur bei Schweden.

Wenn wir eine grössere Anzahl von einander unabhängigen Eigenschaften beobachten, so werden sich die Verteilungskurven von einander verschieben. Nehmen wir an, dass die Variationsbreite nicht grösser ist als 6σ, so wird der Schnittpunkt bei Beobachtung von 9 von einander unabhängigen Eigenschaften auf der Abszisse liegen. Hier haben wir praktisch 100% -ige Sicherheit bei der Individualdiagnose. Bei mehr als 9 Eigenschaften trennen sich die Kurven, einen grösseren Sicherheitsgrad können wir nicht erzielen.

Es dürfte unnötig sein darauf hinzuweisen, dass die hier angewandte Methodik überall da angewendet werden kann, wo man Gruppen verschiedener Art in Bezug auf das Vorhandensein gewisser Eigenschaften von einander unterscheiden will, und dass sie also nicht ausschliesslich von Interesse ist, soweit es sich um Rassenprobleme handelt.

LUIGI L. CAVALLI

Un metodo rapido di calcolo della correlazione media fra più caratteri.

Nello studio della correlazione fra più caratteri è sovente utile il calcolo della media dei coefficienti di correlazione fra gli m caratteri a due a due. Tali coefficienti sono in numero di $m\,(m-1)/2$, e se il numero di caratteri è elevato la massa di calcoli necessari è notevole. Gini [1] ha già messo in evidenza l'importanza della correlazione media e ha descritto un modo abbreviato di calcolo mediante il coefficiente di correlazione sintetico, che dà però la media ponderata dei coefficienti *quadratici* di correlazione fra le variabili a due a due. In un articolo precedente [2] ho indicato un metodo per il calcolo abbreviato della media fra i coefficienti di correlazione *lineari*. La pratica biometrica mi ha convinto però che un metodo più rapido anche se approssimato avrebbe potuto dimostrarsi utile in alcuni tipi di problemi.

Il metodo qui esposto permette una rapidità di calcolo notevole ed entro limiti soddisfacenti dal punto di vista pratico i risultati della determinazione di r non sono affetti da errore sistematico. Il metodo è valido sotto ipotesi di correlazioni lineari fra variabili normali, e d'altra parte i dati biometrici si approssimano soddisfacentemente a questa condizione.

Le informazioni metriche su due o più caratteri in parecchi individui possono essere ridotte semplicemente alla informazione se il carattere i-esimo nell'individuo j-esimo supera o no la media dello stesso carattere nell'intera popolazione. Ad esempio, dati di questo genere :

---

(1) C. Gini, *Metron*, vol. XIV, nn. 2-3-4, 1941.

(2) L. Cavalli, *Metron*, vol. XV, nn. 1-2-3-4, 1949.

		Carattere		
		1	2	3.....
Individuo	1	13	45	40.....
	2	15	63	51.....
	3	17	71	43.....
	4	19	89	62.....
	5	12	70	68.....
	6	10	60	30.....
	7	15	41	28.....
	8	21	46	31.....
	..	..	..	.....
Media		14.5	50.2	41.0...

possono essere trasformati nel modo seguente, limitando l'informazione metrica alla indicazione di *maggiore o minore della media* mediante il segno + o —:

		Carattere.		
		1	2	3
Individuo	1	—	—	—.....
	2	+	+	+.....
	3	+	+	+.....
	4	+	+	+.....
	5	—	+	+.....
	6	—	+	—.....
	7	+	—	—.....
	8	+	—	—.....
	..	..	..	.. ....

Se ne può ricavare un indice, che chiamerò della *concordanza di segno*.

1. Correlazione fra due caratteri.

Per il caso della correlazione di due soli caratteri possiamo supporre di aver già raccolto in una tabella di correlazione i dati singoli. Eccone un esempio [1]:

---

[1] Tratto dal lavoro di Carlifanti e Cavalli, 1945, Boll. I. S. M. 24:215.

Carattere x

Carattere y	0	1	2	3	4	5	6
0	0	13	4	2			
1		10	19	5	1	2	
2		4	19	22	4		
3			5	15	9	3	
4				2	4	3	2
5					2		

Le medie degli x e degli y sono rispettivamente 2,5 e 1,8. Fra il 2 e il 3 per gli x, fra 1 e 2 per gli y tracciamo le due rette (tratteggiate) che dividono la tabella in quattro quadranti.

I segni per x e y sono, in questi quattro quadranti:

x	y	
—	—	nel quadrante in alto a sinistra
—	+	nel quadrante in basso a sinistra
+	—	nel quadrante in alto a destra
+	+	nel quadrante in basso a destra

Ammettendo che le distribuzioni degli x e degli y siano normali (o almeno simmetriche) ed x indipendente da y, la frequenza teorica di ognuno dei quattro quadranti sarà evidentemente $N/4$, se N è il numero totale di individui. Invece dalla tabella abbiamo, riunendo le frequenze contenute in ogni quadrante:

		x —	x +	Totale
y	—	54	10	64
	+	28	66	94
Totale		82	76	158

La frequenza relativa f degli individui con segno concorde (caselle ++ e ——) cioè $(54 + 66)/158$, o la frequenza $g = 1 - f$ degli individui con segno discorde (caselle + — e — +), cioè $(10 + 28)/158$ possono essere prese a misura della correlazione fra x e y. Meglio ancora conviene, facendo uso dei consueti criteri del Gini, stabilire un indice che valga zero per il caso di in-

dipendenza, $\pm$ 1 per i casi di dipendenza completa diretta e inversa. Tenendo conto che per l'indipendenza $f = 0{,}5$, per la dipendenza diretta e inversa $f = 1$ e 0 rispettivamente, si ottiene l'indice :

$$c = 2f - 1 = 1 - 2g \qquad (1)$$

che si può indicare con la lettera c, cioè *coefficiente di concordanza di segno*. Nel nostro caso $f = 120/158 = 0{,}76$, $c = 0{,}52$. Il coefficiente di correlazione lineare r, calcolato su questi dati, vale 0,71.

L'indice c è di calcolo semplice anche senza ricorrere alla tabella di correlazione ; basta in questo caso scrivere al posto di

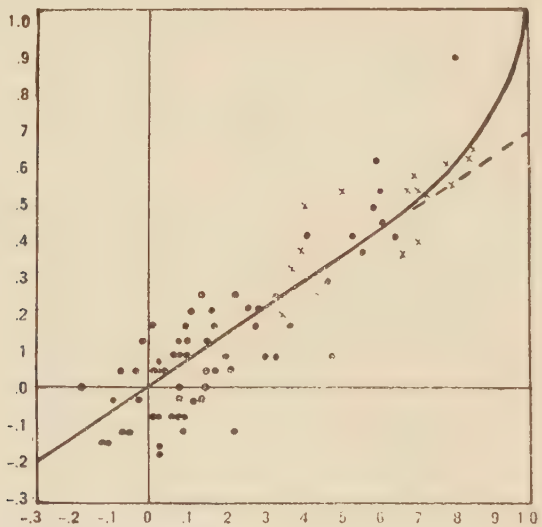


Fig. 1. — Coefficienti di correlazione (in ascissa) e di concordanza (in ordinata) calcolati su materiale zoologico (circoli neri) ed immunologico (croci). I circoli neri si riferiscono a campioni di cinquanta coppie di osservazioni, le croci a campioni di 100-150 coppie di osservazioni.

ogni coppia di valori individuali, un + e un — secondo che il valore individuale è superiore o inferiore alla media per il carattere rispettivo ; indi contare il numero di individui che hanno ugual segno per i due caratteri (cioè gli individui + + e — —), calcolare f dividendo questo numero per il totale degli individui e inserire f nella (1).

Val la pena di notare che può convenire talora riferire le deviazioni + non alla media aritmetica ma alla mediana. Si realizza con ciò anche un risparmio di calcoli.

Perchè questo coefficiente possa divenire utile, occorre :

1) conoscere i rapporti fra c ed r, il coefficiente di correlazione lineare ;

2) conoscere la distribuzione di campionamento casuale di c in funzione di N, numero totale di individui ;

3) estendere a più di due caratteri la misura di concordanza.

2. Rapporti col coefficiente di correlazione lineare r.

Se le variabili x e y sono normali e linearmente correlate, l'elemento di frequenza è :

$$dF = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\, e^{-\frac{1}{2(1-\varrho^2)}\left(\frac{x^2}{\sigma^2 x}-\frac{2\varrho_{xy}}{\sigma_x\sigma_y}+\frac{y^2}{\sigma^2 y}\right)} dx\,dy \qquad (2)$$

indicando con x e y gli scostamenti dalle medie rispettive $\bar{x}$ e $\bar{y}$. La frequenza f del paragrafo precedente sarà evidentemente :

$$f = \int_{-\infty}^{0}\int_{-\infty}^{0} dF + \int_{0}^{+\infty}\int_{0}^{+\infty} dF$$

ovvero anche

$$\frac{f}{2} = \int_{0}^{+\infty}\int_{0}^{+\infty} dF = \int_{0}^{+\infty}\int_{0}^{+\infty} \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\, e^{-\frac{1}{2(1-\varrho^2)}\left(\frac{x^2}{\sigma^2 x}-\frac{2\varrho_{xy}}{\sigma_{xy}}+\frac{y^2}{\sigma_y^2}\right)} dx\,dy \qquad (3)$$

Per risolvere questo integrale [1] poniamo :

$$\frac{x}{\sigma_x} = r\cos\theta \qquad \frac{y}{\sigma_y} = r\,\mathrm{sen}\,\theta$$

---

[1] Desidero ringraziare il dott. C.A.B. Smith e il dott. A. Pratelli che mi hanno molto cortesemente indirizzato nella soluzione di questo integrale.

allora $dx\,dy = r\sigma_x\,\sigma_y\,dr\,d\theta$ e l'integrale (3) diviene:

$$\frac{f}{2} = \int_0^{\pi/2} d\theta \int_0^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}}\, e^{-r^2\left(\frac{1-2\rho\cos\theta\,\mathrm{sen}\,\theta}{2(1-\rho^2)}\right)} r\,dr =$$

$$= \int_0^{\pi/2} \frac{\sqrt{1-\rho^2}}{2\pi}\,\frac{d\theta}{1-2\rho\cos\theta\,\mathrm{sen}\,\theta}$$

Per integrare quest'ultimo poniamo tang $\theta = t$:

$$\frac{f}{2} = \frac{\sqrt{1-\rho^2}}{2\pi}\int_0^{\infty}\frac{dt}{1-2\rho t+t^2} = \frac{\sqrt{1-\rho^2}}{2\pi}\int_0^{\infty}\frac{dt}{(t-\rho^2)+(1-\rho^2)} =$$

$$= \frac{\sqrt{1-\rho^2}}{2\pi}\left[\frac{1}{\sqrt{1-\rho^2}}\,arctg\,\frac{t-\rho}{\sqrt{1-\rho^2}}\right]_0^{\infty} = \frac{1}{2\pi}\left(\frac{\pi}{2} - arctg\frac{-\rho}{\sqrt{1-\rho^2}}\right) =$$

$$= \frac{1}{4} + \frac{1}{2\pi}\,arctg\,\frac{\rho}{\sqrt{1-\rho^2}}$$

e infine ponendo $\rho = \mathrm{sen}\,\varphi$, $\frac{f}{2} = \frac{1}{4} + \frac{\varphi}{2\pi}$

donde, poichè $c = (2f - 1)$:

$$\rho = \mathrm{sen}\left(\frac{\pi}{2}\,c\right) \qquad (4)$$

Per mezzo della (4) è possibile calcolare r in base a c, e chiameremo r_s la stima di r che così si ottiene. Nell'esempio precedente avevamo ottenuto $c = 0{,}52$, che corrisponde a $r_c = 0{,}73$; invece r calcolato in base ai dati metrici vale 0,71.

Nella figura 1 i punti neri corrispondono a valori di r e c calcolati su materiale biometrico (ogni coefficiente calcolato su 50 individui; i coefficienti r sono quelli già pubblicati [1]), i coefficienti c sono dati in tabella 1. La dispersione dei punti attorno alla curva che esprime la relazione teorica fra r e c è notevole

---

[1] L. L. Cavalli, Mem. Ist. Ital. Idrobiol. 2 : 301, 1945.

ma è da tenere presente che la differenza fra r calcolato sui dati metrici e quello calcolato in base a c, cioè r_c (che corrisponde graficamente alla distanza orizzontale fra ogni punto e la curva teorica di fig. 1) non è dovuta solo alla approssimazione di r_c. Infatti la differenza $(r - r_c)$ che ha uno scarto quadratico medio di 0.16 (calcolato sui soli dati di 50 individui, e cioè i punti neri sul grafico) ed esprime l'errore che si commette prendendo, nelle condizioni dette, r_c al posto di r, non dipende solo dall'errore di campionamento casuale $(r_c - \rho)$, ma anche dall'errore di uguale origine $(r - \rho)$, poichè sia r_c che r costituiscono due stime empiriche del valore vero ignoto ρ della popolazione.

È interessante in pratica notare che almeno fino al valore di $\rho = 0.6$, il decorso della funzione $c = f(\rho)$ è quasi lineare, con coefficiente angolare di circa 2/3; per cui si può con buona approssimazione porre:

$$r_c \backsim \frac{3}{2} c \qquad (5)$$

formula che può servire per il calcolo rapido di r_c, per valori di r_c fino a 0,6.

3. Distribuzione di c.

Ha interesse studiare la distribuzione di c per il caso di indipendenza, ed è facile giungervi considerando ancora la tabella tetracorica:

		x —	x +
y	—	a	b
	+	c	d

in cui le frequenze teoriche per distribuzioni normali indipendenti sono $N/4$ in ogni casella.

Il χ^2 che si ottiene per questa ipotesi:

$$\chi^2_{tot.} = \frac{\left(a - \frac{N}{4}\right)^2 + \left(b - \frac{N}{4}\right)^2 + \left(c - \frac{N}{4}\right)^2 + \left(d - \frac{N}{4}\right)^2}{\frac{N}{4}}$$

ha tre gradi di libertà ; ma può venire analizzato nei suoi tre gradi di libertà, ottenendosi tre χ^2, due dei quali esprimono le condizioni di simmetria della x e della y il terzo la condizione di indipendenza di x e y. Ognuno di questi χ^2 ha un solo grado di libertà.

$$\chi^2_x = \frac{\left(a + c - \frac{N}{2}\right)^2 + \left(b + d - \frac{N}{2}\right)^2}{\frac{N}{2}} \text{ simmetria delle } x$$

$$\chi^2_y = \frac{\left(a + b - \frac{N}{2}\right)^2 - \left(c + d - \frac{N}{2}\right)^2}{\frac{N}{2}} \text{ simmetria delle } y$$

$$\chi^2_{xy} = \frac{\left(a - d - \frac{N}{2}\right) + \left(b - c - \frac{N}{2}\right)^2}{\frac{N}{2}} \text{ indipendenza } x - y$$

È facile controllare che :

$$\chi^2_x + \chi^2_y + \chi^2_{xy} = \chi^2_{tot.}$$

La condizione di simmetria espressa dai due primi χ^2 corrisponde alla condizione di uguaglianza della media e mediana.

Si verifica facilmente che :

$$c^2 N = \chi^2_{xy} \tag{6}$$

ove N è il numero totale di individui.

Perciò, nel caso di indipendenza, c^2N è distribuito come un χ^2 con un solo grado di libertà.

Nel caso di dipendenza, poniamo φ = valore vero di f, $\gamma = (2\varphi - 1)$ valore vero di c. Il χ^2 calcolato per un grado di libertà :

$$\chi^2 = \frac{(a + d - \varphi N)^2}{\varphi N} + \frac{[b + c - (1 - \varphi) N]^2}{(1 - \varphi) N}$$

si può facilmente trasformare nel modo seguente :

$$\chi^2 = \frac{(c - \gamma)^2}{1 - \gamma^2} N$$

che dà la distribuzione di c intorno a γ.

4. Correlazione fra m caratteri.

L'interesse di c, più che come metodo rapido approssimato di calcolo di un coefficiente di correlazione lineare, sta nelle possibilità che esso offre di calcolare rapidamente la media di $m\,(m - 1)/2$ coefficienti di correlazione fra m caratteri.

Si consideri ancora la tabella riportata all'inizio del primo paragrafo, in cui n individui esaminati per m caratteri sono stati indicati semplicemente con n serie di m segni + e —, secondo che ognuno degli mn valori individuali superava o no la media per il corrispondente carattere. Si tratta semplicemente, una volta costruita la tabella dei + e dei —, di contare il numero dei + (o quello dei —) di ogni individuo.

Nella tabella già data era :

Individuo N°	1	2	3	4	5	6	7	8
Numero dei +	0	3	3	3	2	1	1	1
Numero dei —	3	0	0	0	1	2	2	2

e considerando per esempio soltanto i +, si possono riunire questi dati in una distribuzione di frequenza :

Numero dei +	0/3	1/3	2/3	3/3	Tot.
Frequenza osserv.	1	3	1	3	8
Frequenza teorica	1	3	3	1	8

Nel caso di indipendenza fra i caratteri si otterrebbe una distribuzione casuale di + e di —, e le frequenze teoriche sarebbero perciò quelle indicate da una serie di Bernoulli in cui la probabilità fondamentale q sia uguale a 1/2. Eseguendo il conteggio su un materiale già oggetto di calcoli precedenti, e dai quali furono ottenuti 66 coefficienti di correlazione (fra 12 caratteri a due a due), ciascuno basato su 50 individui — vedi fig. 1 e tabella 2 — si ottenne :

Numero dei +	0	1	2	3	4	5	6	7	8	9	10	11	12	Tot.
Freq. osserv.	0	2	2	7	5	7	9	3	8	0	5	2	0	50
Freq. teor.	0,0	0,1	0,8	2,7	6,0	9,7	11,4	9,7	6,0	2,7	0,8	0,1	0,0	50,0

mentre le frequenze teoriche per l'indipendenza sono quelle indicate nella riga inferiore. La varianza delle frequenze teoriche (relative) per l'indipendenza è $q(1-q)/m$, ma poichè q, la probabilità fondamentale, è $1/2$, la varianza delle frequenze teoriche è semplicemente $(4\,m)^{-1}$ Invece nel caso di correlazione positiva completa fra tutti i caratteri la varianza sarebbe $1/4$, poichè in tal caso sarebbero occupate solo le due caselle estreme.

Sembra dunque logico calcolare un indice di *concordanza multipla* ricorrendo alla *varianza* della serie osservata, e costruire l'indice in modo che valga + 1 se la dipendenza è completa (tutti i coefficienti di correlazione uguali ad 1) e **0** se si ha completa indipendenza. Tale indice, che indichiamo con c_m, è :

$$c_m = \frac{4\,m\,s^2 - 1}{m - 1} \qquad (9)$$

ove s^2 è la varianza della distribuzione di frequenze osservata. Nel calcolare tale varianza, però, invece che ricorrere alla media empirica del numero di + o di —, conviene ricorrere alla media teorica, $q = 1/2$; nel nostro caso il numero teorico medio di + è $12/2 = 6$. Tale procedimento permette anche di semplificare i calcoli, riunendo insieme le frequenze delle classi equidistanti dal centro. I calcoli si disporranno perciò come nella tabella seguente :

Numero di +	0/12 12/12	1/12 11/12	2/12 10/12	3/12 9/12	4/12 8/12	5/12 7/12	6/12	Totale
Freq. oss. (f)	0	4	7	7	13	10	9	50
$\lvert x\rvert$	6/12	5/12	4/12	3/12	2/12	1/12	0/12	
x^2	36/144	25/144	16/144	9/144	4/144	1/144	0	
fx^2	0	$\frac{100}{144}$	$\frac{112}{144}$	$\frac{63}{144}$	$\frac{52}{144}$	$\frac{100}{144}$	0	$\frac{337}{144}$

La riga indicata con $|x|$ esprime il valore assoluto della frequenza di + divisa per il numero totale di caratteri, e diminuita di $1/2$.

La varianza osservata è (dividendo non per $N - 1$, come si fa generalmente, ma per N) :

$$s^2 = \frac{\Sigma f x^2}{N} = \frac{337}{144} \cdot \frac{1}{50} = 0,0468$$

e c_m, calcolato colla formula (9), vale :

$$c_m = +0,113333\ldots$$

Il *minimo negativo* che c_m può raggiungere non è naturalmente -1 ; si può calcolarlo pensando che solo la (o le due, se m è dispari) caselle centrali siano occupate : esso vale $-1/(m-1)$ se m è pari, e $-1/m$ se m è dispari.

5. Concordanza multipla e concordanza media.

Per il nostro scopo è importante conoscere i rapporti fra c_m il coefficiente di concordanza multiplo ora calcolato, e i singoli $m\,(m-1)/2$ coefficienti di concordanza fra le variabili a due a due. Nella tabella 2 sono riportati i valori di c per 12 caratteri, presi a due a due, e misurati nei 50 individui di *Myxodiaptomus* Lill.. Sono questi i valori corrispondenti alle ordinate dei punti neri nel grafico 1, i cui valori di ascissa (i coefficienti di correlazione lineare) sono già stati pubblicati nel luogo citato.

Tabella 2

Coefficienti di concordanza fra 12 *caratteri di* Mixodiaptomus laciniatus *Lill., a due a due.*

Caratteri.

	3	4	6	7	9	10	11	12	V. 1	V. 2	V. 3
Car. 1	0,20	0,52	0,48	0,04	0,16	—0,08	0,64	0,60	—0,08	0,20	0,04
3		0,20	0,08	—0,04	0,00	—0,16	0,08	—0,04	—0,08	0,04	—0,04
4			0,16	0,20	0,24	0,24	0,40	0,28	0,00	0,12	—0,12
6				0,08	—0,12	0,04	0,36	0,40	—0,04	0,16	0,00
7					0,08	0,24	0,40	0,44	0,08	0,12	—0,20
9						—0,04	—0,12	—0,16	0,04	0,00	—0,08
10							0,04	0,16	0,04	0,08	—0,16
11								0,88	—0,12	0,16	0,16
12									—0,08	0,04	0,12
V⁰ 1										0,00	0,08
V⁰ 2											0,12

La media aritmetica di questi 66 coefficienti è:

$$\bar{c} = \frac{\Sigma\, c}{66} = \frac{7{,}48}{66} = 0{,}11333\ldots$$

e confrontando col valore più sopra ottenuto, $c_m = c$, come è facile dimostrare per via algebrica.

Per semplicità condurrò la dimostrazione su quattro variabili, indicate con x, y, w, z.

Con f_i per i da 1 a 16 sono date le frequenze di individui aventi una delle sedici possibili combinazioni di segni + e — con quattro caratteri $xywz$, come nella tabellina seguente:

	x	y	w	z	
f_1	+	+	+	+	$f_{(4+)} = f_1$
f_2	+	+	+	—	$f_{(3+)} = f_2 + f_3 + f_4 + f_5$
f_3	+	+	—	+	
f_4	+	—	+	+	
f_5	—	+	+	+	
f_6	+	+	—	—	$f_{(2+)} = f_{(2-)} = f_6 + \ldots + f_{11}$
f_7	+	—	+	—	
f_8	—	+	+	—	
f_9	—	+	—	+	
f_{10}	+	—	—	+	
f_{11}	—	—	+	+	
f_{12}	+	—	—	—	$f_{(3-)} = f_{12} + f_{13} + f_{14} + f_{15}$
f_{13}	—	+	—	—	
f_{14}	—	—	+	—	
f_{15}	—	—	—	+	
f_{16}	—	—	—	—	$f_{(4-)} = f_{16}$

Le frequenze siano calcolate in modo che la loro somma sia 1. Inoltre indichiamo con $f_{(4+)}$ la frequenza di individui con quattro segni +; con $f_{(3+)}$ la frequenza di individui con tre segni +, comunque distribuiti fra i quattro caratteri, ecc. e avremo il raggruppamento indicato nel lato destro della tabellina sovrastante.

Per il calcolo di c_m si possono raggruppare ulteriormente le frequenze. Si otterrà, indicando con $f_{(4\pm)}$ la frequenza di indi-

vidui che hanno 4 + o 4 —, ecc.:

$$f_{(4\,\pm)} = f_{(4\,+)} + f_{(4\,-)}$$
$$f_{(3\,\pm)} = f_{(3\,+)} + f_{(3\,-)}$$
$$f_{(2\,\pm)} = f_{(2\,+)} = f_{(2\,-)}$$

La varianza si calcola:

$$s^2 = \left(\frac{1}{2}\right)^2 f_{(4\,\pm)} + \left(\frac{1}{4}\right)^2 f_{(3\,\pm)}$$

e c_m:

$$c_m = \frac{4\,m\,s^2 - 1}{m - 1} = \frac{4\,f_{(4\,\pm)} + f_{(3\,\pm)}}{3} - \frac{1}{3}$$

Ma dalle frequenze delle sedici combinazioni è possibile calcolare anche i singoli coefficienti di concordanza fra le quattro variabili a due a due, c_{ij} che sono in numero di sei. Infatti $c_{ij} = 2\,f_{ij} - 1$, e f_{ij} è la somma di quelle fra le sedici frequenze f_i in cui sia i che j compaiono col segno concordante. Perciò:

$$c_{xy} = 2\,(f_1 + f_2 + f_3 + f_6 + f_{11} + f_{14} + f_{15} + f_{16}) - 1$$
$$c_{xw} = 2\,(f_1 + f_2 + f_4 + f_7 + f_9 + f_{13} + f_{15} + f_{16}) - 1$$
$$c_{xz} = 2\,(f_1 + f_3 + f_4 + f_{10} + f_8 + f_{13} + f_{14} + f_{16}) - 1$$
$$c_{yw} = 2\,(f_1 + f_2 + f_5 + f_8 + f_{10} + f_{12} + f_{15} + f_{16}) - 1$$
$$c_{yz} = 2\,(f_1 + f_3 + f_5 + f_9 + f_7 + f_{12} + f_{14} + f_{16}) - 1$$
$$c_{wz} = 2\,(f_1 + f_4 + f_5 + f_{11} + f_6 + f_{12} + f_{13} + f_{16}) - 1$$

$$\Sigma\, c_{ij} = 2\,\{6\,f_1 + 3\,(f_2 + \ldots + f_5) + 2\,(f_6 + \ldots + f_{11}) + 3\,(f_{12} + \ldots + f_{15}) + 6\,f_{16}\} - 6 = 12\,f_{(4\,\pm)} + 6\,f_{(3\,\pm)} + 4\,f_{(2\,\pm)} - 6$$

Per ottenere c occorre dividere $\Sigma\, c_{ij}$ per $m\,(m-1)/2$, cioè per 6 in questo caso:

$$\bar{c} = \frac{\Sigma\, c_{ij}}{6}$$

Ma notando che la somma delle singole frequenze è 1, si ha

$$f_{(2\,\pm)} = 1 - f_{(4\,\pm)} - f_{(3\,\pm)}$$

che sostituito nella espressione di Σc_{ij} permette di verificare facilmente che :

$$c_m = \bar{c}$$

cioè che il coefficiente di concordanza multiplo è uguale alla media aritmetica dei singoli coefficienti di concordanza fra gli m caratteri a due a due.

Nella pratica interessa però conoscere la media dei coefficienti di correlazione $\bar{r}$, non $\bar{c}$. Si tratta quindi di trasformare $c_m = \bar{c}$ in r, con l'aiuto della formula (4). Questa trasformazione non darà stime affette da errore sistematico sensibile finchè non vi sia, fra gli m $(m - 1)/2$ coefficienti $\bar{r}$, alcuno che superi 0,60 in valore assoluto. Infatti fino a quel valore di $\bar{r}$, $\bar{c}$ ed $\bar{r}$ sono quasi proporzionali, e alla media aritmetica dei valori dell'uno si può sostituire la media aritmetica (trasformata in r_c) dei valori dell'altro. Se alcuni degli $\bar{r}$ superano 0,60, $\bar{r}$ verrà sopraestimato (una stima di $\bar{r}$ per eccesso non porta gravi inconvenienti nei problemi cui accenneremo). Nel nostro esempio, comunque, benchè alcuni degli r siano superiori a 0,60, questo errore per eccesso non si verifica, perchè a $c_m = 0{,}113$ corrisponde $\bar{r}_{cm} = 0{,}18$, mentre il valore di $\bar{r}$ calcolato realmente dalla media dei coefficienti di correlazione lineare vale 0,19.

6. L'uso pratico di c_m nell'esame di popolazioni.

La misura della divergenza fra popolazioni, quando non esista nessun carattere differenziale sicuro, è un difficile compito, in quanto è delicata la scelta fra i numerosissimi indici proposti. Come ho già cercato di dimostrare in un lavoro precedente, l'errore di classificazione è una misura semplice ed opportuna della divergenza biometrica, che dà immediatamente un'idea concreta di tale divergenza, in quanto ne rende evidente il valore tassonomico. D'altra parte, un errore di classificazione in base a misure multiple si può determinare solo sulla base di un certo criterio di classificazione in base a misure multiple ; di tali criteri ne sono possibili parecchi, senza che sia possibile dire a priori — in molti casi — quale sia il migliore. Fissato il criterio di classifica-

zione, l'errore di classificazione, che è poi una misura della transvariazione fra due popolazioni, può essere :

a) *misurato*, partendo da un campione di individui di una o delle due popolazioni, di cui sia nota l'origine, e classificando singolarmente gli individui di questo campione secondo il criterio detto ; si tratta allora soltanto di fare il rapporto fra le classificazioni sbagliate e tutte le classificazioni eseguite, se si vuole ottenere l'errore di classificazione, oppure calcolare qualche altro indice di transvariazione ;

b) *calcolato*, quando si sia determinato un indice di divergenza fra le medie delle due popolazioni per i caratteri in esame e si conosca una formula, che permetta di calcolare in date condizioni l'errore di classificazione, sulla base di semplici costanti statistiche già calcolate.

Il primo metodo, proposto da Gini e Livada [1] è certo il più efficace in quanto evita qualunque errore dovuto alla non normalità delle distribuzioni, non linearità delle correlazioni, ecc. ; ma è anche il più lungo. In pratica converrà ricorrervi quando il criterio di classificazione utilizzato non consenta il calcolo diretto dell'errore di classificazione, o quando le ipotesi che ne sono alla base sul tipo di distribuzioni e di correlazioni non siano soddisfatte.

Dei vari criteri, che permettono il calcolo diretto dell'errore di classificazione, ricorderò qui quello di Dahlberg, che Gini e Livada hanno rielaborato e sviluppato per la misura della transvariazione fra più caratteri, e quello di Zarapkin, che ho recentemente analizzato allo scopo di calcolarne l'errore di classificazione. In ambedue i metodi la formula che permette il calcolo dello errore di classificazione basa non solo su una opportuna misura della distanza biometrica fra le due popolazioni, ma anche sul valore r [2]. Il calcolo di c_m permette così una valutazione rapida dell'errore di classificazione. In questo e simili problemi il metodo può dunque rivelarsi di qualche utilità nella elaborazione di dati biometrici.

---

[1] Soc. Ital. Statistica, Att. VI-VII Riunione pag. 346 (1933).

[2] Con alcune difficoltà di segno messe in luce da Gini e Livada. Il metodo di Dahlberg si può però opportunamente modificare in modo da poter far uso della media aritmetica r.

RIASSUNTO

Si propone un coefficiente di correlazione, detto *coefficiente di concordanza di segno c*, che tiene conto solo dei segni delle deviazioni individuali dalla media dei rispettivi caratteri. Per due soli caratteri tale indice si calcola come $c = 2f - 1$, ove f è la frequenza relativa di individui con segno concorde per i due caratteri; la relazione teorica fra c e il coefficiente di correlazione lineare r è semplice ed è espressa dalla equazione: $r = \text{sen}\left(\frac{\pi}{2} c\right)$. Si discute anche la distribuzione di c.

Per il caso di più di due caratteri si propone un indice di concordanza *multiplo* c_m che si calcola con altrettanta facilità dalla varianza della distribuzione degli individui con 0, 1, 2,... m segni positivi (o negativi) per gli m caratteri, e che si dimostra essere uguale alla media aritmetica fra gli $m(m-1)/2$ coefficienti di concordanza fra le variabili a due a due.

Si discutono le possibilità di applicazioni pratiche di questi indici di correlazione semplificati, specie in problemi biometrici.

SUMMARY

A simple correlation coefficient, c (coefficient of concordance) is proposed, which takes into account only the signs of the individual deviations from the means of the respective characters. For two characters this coefficient is calculated as $c = 2f - 1$ where f is the relative frequency of individuals with equal (both positive or negative) signs for the two characters. The theoretical relation between c and the linear correlation coefficient r is simple and is given by: $r = \text{sen}\left(\frac{\pi}{2} c\right)$. The sampling distribution of c is discussed.

For the case of more than two characters a multiple index of concordance is proposed, c_m, which can be calculated still very simply from the variance of the distribution of individuals with 0, 1, 2,... m positive (or negative) signs for the m characters, and which is shown to be equal to the arithmetic mean of the $m(m-1)/2$ concordance coefficients between the m characters two by two.

Possibilities of practical application of these simplified correlation coefficients are discussed, especially in relation to biometric problems.

ZUSAMMENFASSUNG

Ein einfacher Korrelationskoeffizient (Konkordanzkoeffizient) c wird vorgeschlagen, der nur auf Grund des Zeichens der individuellen Abweichungen vom mittleren Wert der bez. Charakteren berechnet wird. Für zwei Charakteren dieser Koeffizient wird berechnet als $c = 2f - 1$, wo f ist die Häufigkeit von Individuen mit gleichem Zeichen für beiden Charakteren. Die theoretische Beziehung zwischen c und den gewohnlichen linearen Korrelationskoeffizient r ist einfach : $r = \text{sen}\left(\frac{\pi}{2}c\right)$. Die Verteilung von c wird besprochen.

Für den Fall der Korrelation zwischen mehr als zwei Charakteren ein mehrfacher Korrelationskoeffizient c_m wird vorgeschlagen, der kann einfach aus der Streuung der Verteilung von Individuen mit 0, 1, 2,... m positive (oder negative) Zeichen für die m Charakteren berechnet werden. Es wird bewiesen, dass dieser mehrfache Koeffizient gleich zu dem arithmetischen Mittel sämtlicher $m(m-1)/2$ Konkordanzkoeffizienten zwischen den Charakteren zwei bei zwei ist.

Die Möglichkeiten praktischer Anwendungen dieser vereinfachten Korrelationskoeffizienten, insbesondere bei biometrischen Problemen, werden besprochen.

V. CASTELLANO

Sui quozienti di eliminazione [1]

SOMMARIO: § 1. – Che cosa debba intendersi per «specificazione» dei rapporti di derivazione generici. I rapporti di derivazione come rapporti di composizione. Il rapporto di probabilità. — § 2. – Probabilità di una caratteristica α in una popolazione non omogenea. Metodo dei gruppi scelti. Eliminazione della influenza della frequenza dei vari gruppi scelti, mediante il metodo della «funzione di sopravvivenza» e della «popolazione stazionaria». Indipendenza del metodo della funzione di sopravvivenza dall'ordine di successione dei gruppi scelti. Caso in cui i gruppi scelti costituiscono una successione. Caso in cui costituiscono una serie sconnessa: esempio, e preferenza da accordarsi alla media aritmetica semplice dei quozienti dei gruppi scelti. — § 3. – Popolazioni costituenti collettività aperte. Caso particolare in cui una collettività aperta equivale a una collettività chiusa. Necessità di considerare la legge temporale di eliminazione per la causa α. Tasso istantaneo di eliminazione. Probabilità indipendente e tasso centrale di eliminazione nel caso $k(t) = 0$ (differenza tra entrati e usciti per causa $\neq \alpha$ costantemente nulla). Caso generale $k(t) \neq 0$. Formula classica per la determinazione della probabilità indipendente di eliminazione. Suo particolare campo di applicazione e ipotesi supplementare che richiede la sua pratica applicazione. Formule equivalenti, la cui applicazione non richiede nessuna altra ipotesi. Esempio di calcolo. Caso generale di una collettività aggregata. Il saggio medio, o probabilità media di eliminazione. — § 4. – Perchè il saggio medio può essere > 1. Espressione della probabilità indipendente in funzione del saggio medio. Caso particolare di un tasso istantaneo costante: formula del Gini e formula del Wittstein. Impossibilità di dedurre la formula del Gini dalla formula classica, valida in un campo più ristretto. Ammissibilità della ipotesi $\mu(t) = \text{cost.}$ in una collettività aggregata. Il tasso istantaneo come utile finzione matematica. Legittimità della definizione $\mu(t) = s(t)$, in ogni intervallo di tempo sufficientemente breve. Formula pratica per il calcolo della probabilità indipendente di eliminazione nel caso generale, e formula approssimata del Gini. — § 5. – Determinazione della proba-

[1] Comunicazione presentata il 28 maggio 1946 alla riunione del Seminario di Statistica dell'Istituto di Statistica della Università di Roma.

bilità di un concepimento per anno di età delle madri, in due gruppi di famiglie nobili, riportate nell'Almanacco di Gotha. I due gruppi costituiscono due collettività aggregate : probabilità di concepire nello stesso anno di età nel quale è stato celebrato il matrimonio. Formule per l'applicazione dei metodi delle durate esatte dell'Ackland e delle durate esatte del Gini. Confronto con le probabilità annue dedotte dalle medie aritmetiche, semplice e ponderata dei saggi medi mensili di eliminazione ottenuti col metodo delle durate medie del Gini.

§ 1. — Il quoziente di natalità è un quoziente di derivazione in quanto mette in relazione la intensità di un fenomeno (nati) con quella di un altro fenomeno (popolazione) che ne costituisce il presupposto generico ; ma poichè nella popolazione sono distinguibili chiaramente vaste categorie d'individui che a nascite non possono dar luogo – vecchi e bambini – esso viene meglio specificato, depurando la popolazione di tutte quelle classi che un più approfondito esame dimostra non poter dar luogo a nascite, fino a istituire il rapporto tra il numero dei nati e quella parte della popolazione atta alla generazione, per es. : donne in età feconda, nella quale non si può cercare di approssimare meglio la più ristretta parte che produrrà effettivamente tutte le nascite di una certa unità di tempo, senza correre l'alea di escludere anche una frazione, sia pure piccola, di dette nascite.

Il quoziente di mortalità è un quoziente non suscettibile di maggiore specificazione, perchè non esiste individuo che non possa, a priori, morire in un certo tempo assegnato.

Sia per l'uno che per l'altro rapporto pure si calcolano valori relativi alle singole classi di età della popolazione, che vanno sotto il nome, ormai consacrato dall'uso, di quozienti specifici, generando naturalmente il dubbio che si tratti di una ulteriore specificazione dei quozienti stessi, mentre non si tratta di una ulteriore riduzione della popolazione che si considera, ma del suo frazionamento in classi, secondo le modalità di un certo carattere (età), in ciascuna delle quali il quoziente in oggetto assume un particolare valore.

Se per processo di specificazione di un quoziente $A \mid B$ s'intende quel procedimento mediante il quale nella massa dei casi del fenomeno B – presupposto generico del fenomeno A – si delimita progressivamente la più ristretta delle collettività contenute in B, dalla quale può ancora trarre origine la collettività A, al-

lora i quozienti di natalità, nuzialità, mortalità e simili, per classi di età, non hanno diritto di chiamarsi tutti allo stesso modo, specifici: se invece per specificazione volesse intendersi la progressiva scomposizione della collettività *B* in gruppi scelti che differiscono sistematicamente nei riguardi del fenomeno *A*, detti coefficienti per classi di età non possono che segnare una tappa più avanti nel malcerto cammino della classificazione, e quindi anche in questo caso l'attributo di «specifici» deve considerarsi errato nella maggior parte dei casi nei quali non sia stata verificata la omogeneità degl'individui di ciascuna classe rispetto alla caratteristica che si esamina. E, comunque, tale omogeneità è sempre più supposta che reale. Per es., si assumono come gruppi omogenei, rispetto alla possibilità di concepire, le coniugate di uguale età e uguale durata di matrimonio, ma, in effetti, non tutte concepiscono colla stessa facilità, e alcune, per malattia o cattiva conformazione, costituiscono addirittura casi impossibili. Il quoziente «specifico» starebbe in ogni caso a indicare che si è raggiunto il grado di specificazione oltre il quale non si può andare senza procedere ad una discriminazione di casi individuali alla cui impossibilità o estrema complessità si deve appunto la trattazione dei problemi «con metodo statistico».

Poichè i progressi di una disciplina sono certamente agevolati dalla precisione del linguaggio che essa si crea, tale terminologia dovrebbe essere opportunamente precisata. Attenendomi al primo concetto di «specificazione», a me pare siano da indicare con l'attributo di «specifici» tutti quei quozienti di derivazione nei quali la collettività primitiva non si può ulteriormente ridurre senza frazionare la collettività derivata: così tutti i quozienti di mortalità, sia relativi alla popolazione generale che a gruppi comunque scelti, sono specifici; mentre non sono quozienti specifici i quozienti di natalità per età, e così via.

I quozienti di derivazione sono sempre assimilabili a rapporti di probabilità, in quanto esprimono la frequenza, constatata alla fine di un certo intervallo di tempo, di un certo fenomeno *A* (morte, infortunio, ecc.,) tra i casi di un certo fenomeno *B* (popolazione). Il fatto che, una volta definita la popolazione, tale frequenza non sia constatabile che alla fine di un certo intervallo di tempo, non costituisce che una variante al concetto classico di probabilità come rapporto di composizione, che è quello

del Gini (1), variante però legittima, nulla impedendo di considerare da un punto di vista formale, che i singoli individui che effettivamente morranno nell'intervallo di tempo T sieno già all'inizio differenziati dagli altri, ma in maniera non osservabile da noi. La popolazione è pensabile quindi scomposta, rispetto al carattere « morte», in due classi d'individui : « che muoiono » e « che non muoiono » nel tempo T, come le palline, nel classico esempio dell'urna, rispetto al colore, in nere e bianche. Ciò equivale d'altra parte, a concepire i rapporti di derivazione come rapporti di composizione, il che è sempre possibile, sotto condizioni molto late, quando si abbia riguardo a una collettività chiusa, la cui composizione cioè si possa immaginare costante durante il tempo, sufficientemente breve, della osservazione.

§ 2. — La considerazione di gruppi scelti nella popolazione – col quale ultimo nome possiamo designare in generale una qualsiasi collettività di casi, i quali ne costituiscono gl'individui – porta a rappresentare il quoziente di derivazione – e in partico-

(1) È la definizione del Gini : probabilità = frequenza totalitaria (cfr. meglio che : *Che cosa è la probabilità ?* in : « Scientia », vol. III, A. II, 1908 ; *Alle basi del metodo statistico. Il principio della composizione degli errori accidentali e la legge dei grandi numeri*, in « Metron », vol. XIV, n. 2-4, 1941, pagg. 225-228, dove il concetto è espresso con la maggiore efficacia) non differisce, in sostanza, dalla definizione classica, che per una maggiore concretezza, indispensabile ai concetti matematici per trovare posto nella Statistica. Infatti, per probabilità di una pallina bianca, nel classico esempio dell'urna, non si definisce la frequenza delle palline bianche nell'urna ? Ma alla definizione classica si aggiunge ordinariamente una coda : « che tutti i casi possibili siano ugualmente possibili, o probabili » e questa coda costituisce, col corpo della definizione, un bisticcio piuttosto oscuro, risolto d'ordinario colla condizione della identicità fisica dei vari casi possibili, come se non si potesse giocare equamente a testa croce, con una moneta imperfetta, o al gioco del lotto se i numeri fossero scritti sopra sassolini qualunque invece che su belle sferette di avorio o di legno. Mi pare questo il più bell'esempio di disgrazia toccata a un concetto di per sé compiuto e chiarisimo, per insufficienza di linguaggio. Infatti, quella coda, non ha altro ufficio che quello di permettere di constatare che la probabilità è l'aspettativa della frequenza nei campioni. E i campioni si compongono di osservazioni, che debbono poter cadere indifferentemente sull'uno o sull'altro dei casi osservabili. Non si tratta quindi d'identità fisica, ma di uguale osservabilità dei casi possibili, e questa può essere realizzata benissimo anche se i singoli casi sono materialmente dissimili.

lare il rapporto di probabilità, al quale faremo spesso espresso riferimento – come media ponderata dei rapporti relativi ai vari gruppi scelti, essendo pesi i numeri degl'individui dei singoli gruppi, cioè: la probabilità di una caratteristica α in una popolazione P risulta espressa da

$$ {}_{\alpha}q_{p} = \frac{1}{n} \sum n_i \; {}_{\alpha}q_{pi} \qquad /1/ $$

dove coll'indice $i = 1,2,..,s$ sono contrassegnati i vari gruppi scelti P_i nel proprio ordine, se presentano un ordine naturale di successione, come ad es., gli anni di età, o in un ordine qualsiasi, se un ordine naturale non esiste, come nel caso delle modalità dello stato civile, dei sessi, delle professioni, ecc. Sopprimendo la indicazione della popolazione P, quando non dia luogo ad equivoci, e indicando con « caratteristica α_i » la caratteristica α nel gruppo scelto i.esimo, la /1/ si scrive:

$ {}_{\alpha}q = \frac{1}{n} \sum n_i \; {}_{\alpha_i}q $ /1'/, e sopprimendo anche la indicazione della caratteristica, ove non sia possibilità di equivoco:

$$ q = \frac{1}{n} \sum n_i \; q_i \,. \qquad /1''/ $$

Applichiamo successivamente i quozienti ${}_{\alpha_i}q$ ad una collettività di l_0 individui, nella quale dunque si suppone che la caratteristica α si presenti colla frequenza ${}_{\alpha_1}q$ in una prima unità di tempo, ${}_{\alpha_2}q$ nella seconda, e così via. Il numero degl'individui presentanti la caratteristica α nelle successive unità di tempo sarà: $l_0 \; {}_{\alpha_1}q$; $l_0 (1 - {}_{\alpha_1}q) \cdot {}_{\alpha_2}q$; $l_0 (1 - {}_{\alpha_1}q) \cdot (1 - {}_{\alpha_2}q) \cdot {}_{\alpha_3}q$; e così via, sicché avremo un numero complessivo d'individui contrassegnati colla caratteristica $\alpha - o$ più brevemente, « individui α », come li indicheremo nel seguito: $l_0 \; {}_{\alpha_1}q$ al tempo 1; $l_0 \{ 1 - (1 - {}_{\alpha_1}q) (1 - {}_{\alpha_2}q) \}$ al tempo 2; $l_0 \{ 1 - (1 - {}_{\alpha_1}q) \; (1 - {}_{\alpha_2}q) \; (1 - {}_{\alpha_3}q) \}$ al tempo 3, e così via. A procedimento ultimato, la frequenza ${}_{\alpha_1,2,\cdot,s}q$ degl'individui α nel gruppo degli l_0 esposti risulterà:

$$ {}_{\alpha_1,2,\ldots,s}q = 1 - (1 - {}_{\alpha_1}q) (1 - {}_{\alpha_2}q) \ldots (1 - {}_{\alpha_s}q) \qquad /2/ $$

nota formula del calcolo delle probabilità, che esprime in questo

caso la probabilità che abbia a presentare la caratteristica α un individuo che entri successivamente a far parte dei vari gruppi scelti P_i, non importa in quale ordine.

Ponendo $1 - {}_{\alpha_i}q = {}_{\alpha_i}p$, e sostituendo p_i a ${}_{\alpha_i}p$, la /2/ si scrive più semplicemente:

$$q = 1 - p_1\, p_2 \ldots p_s \qquad /2'/$$

Ogni particolare ordine di successione dei gruppi scelti P_i definisce una legge di eliminazione nella collettività l_o, con la quale si associa a ciascun quoziente ${}_{\alpha_i}q$ il peso costituito dal numero

$$l_{i-1} = l_o\,(1 - {}_{\alpha_1,\,2,\,\ldots\,,i-1}q) \qquad /3/$$

dei non eliminati per le cause $\alpha_1, \alpha_2, \ldots \alpha_{i-1}$, cioè non presentanti ancora la caratteristica α. Il numero degl'individui α derivati, nelle successive unità di tempo, da l_o, può anche riguardarsi come il numero degl'individui α che si producono, nella unità di tempo, in una popolazione nella quale le cause α_i operano su gruppi scelti di l_i individui. Tale popolazione è la popolazione che genererebbe la legge di eliminazione assegnata, operando, ad ogni unità di tempo, su un nuovo contingente di l_o individui, e su ciascuno, per s unità di tempo. In questa popolazione la frequenza degli individui α risulta, nella unità di tempo:

$${}_{\alpha}\bar{q} = \sum_{1}^{s} l_{i-1}\ {}_{\alpha}q_i \,/\, \sum_{1}^{s} l_{i-1} \qquad /4/$$

mentre nella popolazione degli l_o tale gruppo d'individui presenta la frequenza:

$${}_{\alpha_1,\,2,\,\ldots\,,s}q = \sum_{1}^{s} l_{i-1}\ {}_{\alpha_i}q \,/\, l_o \qquad /5/$$

che evidentemente coincide con la /2/. Tra la /4/ e la /5/ passa la relazione

$${}_{\alpha}\bar{q} = \frac{l_o}{\sum_{o}^{s-1} l_i}\ {}_{\alpha_1,\,2,\,\ldots\,,s}q \qquad /6/$$

e da questa si ricava, per la (4) la espressione /4'/:

$${}_{\alpha}\bar{q} = \frac{1 - p_1\, p_2 \ldots p_s}{\frac{1}{l_o}\ \sum_{o}^{s-1} l_i} \qquad /4'/$$

Nella legge di eliminazione rappresentata dai $\alpha_i q$ nell'ordine, gli eliminati da l_o nella i.esima unità di tempo, vivono, ciascuno, non meno di $i - 1$ e non più di i unità di tempo. Il numero delle unità di tempo quindi globalmente trascorse per gli eliminati, prima del verificarsi dell'evento α, aumentato delle $s\, l_s$ unità di tempo totalizzate dai non eliminati in tutto il periodo, soddisfa pertanto alla disuguaglianza:

$$\sum_{1}^{s} (i-1)\,(l_{i-1} - l_i) + s\, l_s \leq H \leq \sum_{1}^{s} i\,(l_{i-1} - l_i) + s\, l_s\,,$$

cioè:

$$\sum_{o}^{s-1} l_i - (l_o - l_s) \leq H \leq \sum_{o}^{s-1} l_i\,. \qquad /7/$$

Il rapporto H / l_o indica la vita media degl'individui di l_o. Dalla /7/ segue che:

$$\sum_{o}^{s-1} (l_i / l_o) - (H / l_o) \leq (l_o - l_s) / l_o = \alpha_{1,2,\ldots,s} q.$$

Indicata con θ una opportuna quantità ≤ 1, si ha:

$$\sum_{o}^{s-1} l_i / l_o = (H / l_o) + \theta\, \alpha_{1,2,\ldots,s} q.$$

$\sum_{o}^{s-1} l_i / l_o$ è pertanto un valore approssivmato per eccesso $\bar{v}_m$ della vita media $v_m = H / l_o$ degli individui di l_o. Scriveremo quindi la /6/ anche

$$\alpha \bar{q} = \alpha_{1,2,\ldots,s} q / \bar{v}_m\,. \qquad /6'/$$

La /5/ si ricollega alla formula nota in matematica attuariale, specialmente sotto la forma continua, col nome di probabilità indipendente, o pura, o assoluta, o parziale, di J. Karup (1). Tale lusso di aggettivi può generare, in qualche caso, confusione, e per questo mi atterrò, in generale al primo. Il suo campo di applicazione è tuttavia più ampio di quello della probabilità indipendente di Karup, in quanto, nel campo discontinuo, essa si stabilisce senza far ricorso alla ipotesi di un ordine naturale nella successione delle classi p_i, risultando la /5/ invariante per tutte le permutazioni dell'indice i, come mostra chiaramente la /2/.

(1) Cfr.: F. Cantelli, *Genesi e costruzione delle tavole di mutualità*, Ministero di Agricoltura, Industria e Commercio, Roma, 1914.

Considerare i gruppi scelti in un certo ordine o in un ordine diverso, non influisce evidentemente sul calcolo della frequenza della caratteristica α nella popolazione reale, quindi non muta neppure il valore $_{\alpha}q$ espresso dalla /1/ o /1'/. Mutando invece il sistema l_i dei pesi, l'ordine dei $_{\alpha_i}q$ influisce sul valore di $_{\alpha}\bar{q}$, espresso dalla /4/. Le $s!$ leggi di eliminazioni che si ottengono considerando tutte le possibili permutazioni dei $_{\alpha_i}q$ producono $s!$ popolazioni stazionarie collegate, le quali, pur fornendo lo stesso numero d'individui α nella unità di tempo, differiscono l'una dall'altra per il numero complessivo degl'individui. A seconda dell'ordine nel quale si considerano i gruppi scelti la /4/ fornirà per $_{\alpha}\bar{q}$ s valori diversi. In base a un teorema generale sulle medie ponderate (1) i valori minimo e massimo di $_{\alpha}\bar{q}$ si otterranno ordinando i gruppi secondo i valori crescenti e decrescenti dei $_{\alpha_i}q$, perchè, essendo la successione dei pesi l_i decerscente, si ottenga, nel 1° caso, contrograduazione e, nel 2° cograduazione, tra pesi l_i e intensità $_{\alpha}q_i$. Reciprocamente, la vita media $\bar{v}_m$, legata alla $_{\alpha}\bar{q}$ dalla /6'/ assumerà, nei due casi ora detti, il suo massimo e il suo minimo valore.

La /6/ (o la /6'/), che nelle pratiche applicazioni può assumere la forma semplicissima

$$_{\alpha}q = (l_o - l_s) \; / \; \sum_{o}^{s-1} l_i \qquad /6''/$$

risolve soddisfacentemente il problema del confronto di due successioni di quozienti di eliminazione, in tutti quei casi nei quali si fa ricorso ad una più o meno arbitraria popolazione tipo.

Sulla insufficienza teorica del metodo della popolazione tipo si è soffermato di recente, il Galvani, e, del resto, il metodo, insieme al favore degli statistici, si guadagnò critiche, tra tutte decisiva quella sulla dipendenza dei risultati del confronto dalla popolazione tipo adottata (2). Se esiste una particolare popola-

---

(1) Cfr.: V. Castellano.

(2) C. Gini: *Quelques considérations au sujet de la construction des nombres indices des prix et des questions analogues, n.* 41, pagg. 123-28, in: «Metron», vol. IV, 1924; e L. Galvani, *Insufficienza teorica del metodo della popolazione tipo*, in: «Atti della VI riunione scientifica della Società Italiana di Statistica», Roma, 1943. Le obbiezioni che entrambi gli AA. espongono nei riguardi del metodo della popolazione tipo riguardano anche le applicazioni alla determinazione dei quozienti di derivazione, ma

zione rispetto alla quale si debbano, o si vogliano, confrontare le due successioni di quozienti, la questione del metodo non si pone neppure: è ovvio che se interessa il confronto nei riguardi di una certa popolazione, il confronto si esegua rispetto a questa e non ad un'altra; ma se una tale popolazione non esiste e lo statistico deve prodigarsi in sforzi di buona volontà per puntellare da tutte le parti l'edificio traballante delle ragioni che dovrebbero dimostrare la necessità dell'una o dell'altra popolazione tipo – e non dimostrano se non l'assenza di ogni valido argomento – in prò dell'una e dell'altra – allora sarebbe ragionevole, penso,

---

il caso dei numeri indici dei prezzi, da essi considerato in particolare, non rientra se non in una generalizzazione di quest'ultimo problema, che porta a scrivere la /1/ sotto la forma: $\bar{q} = \Sigma\, \alpha_i q\, f_i\, (\alpha_i\, q\,,\, \gamma)\ /\ \Sigma\, f_i\, (\alpha_i q\,,\, \gamma)$ /1''/ colla quale si esprime, nel caso di un quoziente di eliminazione, ad es., per morte, che il totale degli eliminati è la somma degli eliminati di ciascuna classe, l'ammontare f_i di queste dipendendo, e dai quozienti stessi di eliminazione delle varie classi, e dal complesso γ delle altre circostanze che influiscono sulla dinamica della popolazione. Considerati costanti i ${}_{\alpha_i}q$, ogni sistema degli f_i (ogni popolazione tipo) equivale ad una ipotesi sul complesso γ delle altre condizioni dalle quali gli f_i dipendono. Così la /4/ si fonda sulla ipotesi di una popolazione stazionaria. (l_o = cost.). Nel caso del prezzo medio, fissati i prezzi q_i, il prezzo medio q dipende dal sistema delle quantità scambiate f_i, e queste dal complesso di tutte le restanti condizioni di mercato che determinano, insieme ai prezzi, le quantità scambiate. Nelle condizioni reali di mercato, ai prezzi praticati corrispondono le quantità effettivamente scambiate, ma ad ogni ipotesi sulle condizioni di mercato corrisponderà un certo sistema di quantità f_i che varierà al variare dei prezzi q_i, qualora si mantenga poi costante il complesso delle altre circostanze (γ). Questa impostazione del problema mette in chiara luce la illogicità del ricorso al metodo della popolazione tipo (quantità variabili per pesi costanti) nel caso che esista una relazione necessaria tra quantità e pesi, come nel caso della costruzione dei numeri indici dei prezzi (quantità = prezzo, peso = quantità scambiate). Tale illogicità dà un carattere di approssimatività ai numeri indici dei prezzi che il metodo della popolazione tipo porta a costruire, come rileva il Gini (pagg. 80-81), e può darsi che, in periodi normali, risulti in pratica un'approssimazione sufficiente. Ma in vista dell'applicazione anche a periodi anormali, nei quali più fortemente variano e prezzi e quantità, e nei quali maggiormente si sente il bisogno di numeri indici di chiaro significato, metterebbe conto, credo, di abbordare il problema della costruzione dei numeri indici dei prezzi da un punto di vista più logico, che può riassumersi così: per ogni sistema di prezzi le quantità corrispondenti, compatibilmente colle ipotesi fatte su γ.

adottare per ciascuna successione la propria popolazione tipo, che è la popolazione stazionaria che si formerebbe sotto la sola azione della causa α, solo che la legge di eliminazione operasse su contingenti costanti di nuovi esposti, ad ogni unità di tempo. Milita in favore di ciò anche quel « metodo del minimo arbitrio » introdotto dal Gini nella metodologia, una decina di anni fa (1).

Nel caso classico della mortalità, la (6') assume una forma anche più espressiva, se i quozienti ${}_{\alpha_i}q$ si riferiscono a tutte le classi di età della popolazione. In tal caso, infatti, risulta ${}_{\alpha_1, \alpha_2, \ldots, \alpha_s}q = 1$, in quanto la legge di eliminazione abbraccia tutta la durata della generazione. Risulta pertanto:

$$_{\alpha}\bar{q} = 1 / \bar{v}_m \qquad /8/$$

Se, come spesso, si calcola per v_m il valore della vita media completa $\overset{o}{e} = \sum_{o}^{\omega} l_i / l_o - 1/2$, allora la /8/ diventa:

$$_{\alpha}\bar{q} = 1 / (e^o + 1/2) \quad (2)$$

Ma quando i gruppi scelti della popolazione costituiscano un insieme sconnesso, allora esistono $s!$ popolazioni tipo, collegate con altrettante possibili successioni dei ${}_{\alpha_i}q$, e altrettanti valori

---

(1) C. Gini: *Sulla determinazione della frequenza dei geni dei gruppi sanguigni*, « Genus », vol. 1, n. 3-4, Anno 1935.

(2) Questa formula, sotto la forma $q = \frac{1}{1 + p_1 + p_1 p_2 + \ldots}$ è usata dal Mortara (*Le popolazioni delle grandi città italiane*, in: « Biblioteca dell'Economista », vol. XIX, Torino, 1908), il quale però osserva che nella formula hanno importanza via via decrescente le successive probabilità di sopravvivenza, e quindi i quozienti di mortalità delle età più elevate. Il che non vedo come possa costituire un difetto. Nella pagina seguente, il Mortara riporta il calcolo della vita media, ma non collega i due indici, che, in fondo, sono la stessa cosa. Del resto, la vita media è riconosciuta come un indice corretto della mortalità, da molti AA. Cfr., ad es. il Galvani, *Diminuzione della mortalità in alcuni Stati*, in: « Atti del Congresso internazionale per gli studi sulla popolazione », Roma, 1931, pubblicati dal Comitato Italiano per lo studio dei problemi della popolazione, che adopera tali indici per i confronti nel tempo e nello spazio. Molto esplicito in argomento è il Boldrini (Demografia, Giuffrè, Milano, 1943, pagg. 183-190).

può assumere la $_{\alpha}\bar{q}$, compresi fra un minimo e un massimo alquanto più ravvicinati del minimo e del massimo compatibili, a priori, con una scelta arbitraria del tipo della popolazione [1], ma resta pur sempre, per q, un certo campo di variabilità, che può rendere aleatori i confronti a meno che non si assuma una media di tali valori di $_{\alpha}\bar{q}$. Assumiamo una media ponderata, con pesi uguali ai relativi denominatori. Risulta:

$$_{\alpha}\bar{q} = \frac{l_o\,[1 - (1 - q_1)(1 - q_2)\ldots(1 - q_s)]}{\frac{1}{s!}\,\Sigma \sum_{o}^{s-1} l_i}$$

la prima somma a denominatore essendo estesa a tutte le permutazioni degl'indici i. Ponendo $1 - q_i = p_i$, e indicando con $p_{i]}$ il valore di p_i che occupa l'i. esimo posto nella permutazione generica dei p_i, poichè $l_i = l_o\, p_1\, p_2 \ldots p_i$, segue che:

$$_{\alpha}\bar{q} = s!\ \frac{1 - p_1\, p_2 \ldots p_s}{s! + \sum_{1}^{s-1} \Sigma p_{1]}\, p_{2]} \ldots p_{i]}}$$

La somma $\Sigma p_{1]}\, p_{2]} \ldots p_{i]}$ è estesa agli $s!$ l_i relativi alle $s!$ permutazioni degl'indici i, ma poichè i termini $p_1\, p_2 \ldots p_i$ distinti sono solo in numero di $\binom{s}{i}$, a ciascuno di essi spetterà un coefficiente $s! / \binom{s}{i}$. Sarà pertanto:

$$_{\alpha}\bar{q} = \frac{1 - p_1\, p_2 \ldots p_s}{1 + \sum_{1}^{s-1} 1 / \binom{s}{i} \Sigma p_{1]}\, p_{2]} \ldots p_{i]}} \qquad /9/$$

ove la Σ è estesa a tutte le combinazioni i a i degli indici $1,2,..,s$.

La /9/ è più complicata a scriversi che ad applicarsi. Per $s = 2$ fornisce:

$$_{\alpha}\bar{q} = \frac{1 - p_1\, p_2}{1 + 1/2\,(p_1 + p_2)}. \quad \text{Per } s = 3:$$

---

[1] Cfr. L. Galvani, *Insufficienza teorica ecc.*, già cit.

$$_{\alpha}\bar{q} = \frac{1 - p_1\, p_2\, p_3}{1 + {}^1/_3\,(p_1 + p_2 + p_3) + {}^1/_3\,(p_1\, p_2 + p_1\, p_3 + p_2\, p_3)} =$$

$$= \frac{1 - p_1\, p_2\, p_3}{1 + {}^1/_3 \sum p_i + {}^1/_3 \sum p_i\, p_j}$$

e così via, risultando al denominatore la somma delle combinazioni 0 a 0, 1 a 1, 2 a 2, e così via, delle probabilità di sopravvivenza delle singole classi, affette, nell'ordine, da coefficienti uguali ai reciproci dei coefficienti binomiali.

O può servire, in tali casi, fare ricorso alla /5/, che è indipendente dell'ordine di successione dei $_{\alpha i}q$?

Precisiamo con un esempio. Indichi α_1 l'evento morte nel sesso maschile e α_2 l'evento morte nel sesso femminile. In una popolazione P i due quozienti sono $_{\alpha_1}q$ e $_{\alpha_2}q$, e in un'altra P' sono $_{\alpha_1}q'$ e $_{\alpha_2}q'$. In quale popolazione la mortalità è maggiore, a prescindere dalla influenza della diversa composizione della popolazione per sesso ? Prima di tentare di rispondere a questa domanda, chiediamoci se ha un senso. Che cosa significa eliminare l'influenza della diversa composizione per sesso ? Supporre le popolazioni ugualmente composte ? Il metodo della popolazione tipo risponde a meraviglia. E che la intensità della disuguaglianza risultante tra i due quozienti generali $M\,(_{\alpha}q)$ e $M\,(_{\alpha}q')$ dipenda dalla composizione della popolazione prescelta non può essere meraviglia per nessuno, ché sanno anche gli osti che il prezzo unitario di un miscuglio di vini dipende dalla proporzione dei componenti, se questi hanno prezzi diversi. E se le due disuguaglianze $_{\alpha_1}q \neq {}_{\alpha_1}q'$ e $_{\alpha_2}q \neq {}_{\alpha_2}q'$ hanno lo stesso verso, uguale verso avrà la disuguaglianza $M\,(_{\alpha}q) \neq M\,(_{\alpha}q')$, mentre se hanno versi discordi, potrà accadere e che sia $M\,(_{\alpha}q) > M\,(_{\alpha}q')$, e che sia $M\,(_{\alpha}q) < M\,(_{\alpha}q')$ [1]. È indispensabile precisare la popolazione di confronto, e se non è ragione di preferenza, si assuma

$$M\,(_{\alpha}q) = {}^1/_2\,(_{\alpha_1}q + {}_{\alpha_2}q),\quad M\,(_{\alpha}q') = {}^1/\ \ (_{\alpha_1}q' + {}_{\alpha_2}q')$$

---

[1] Cfr. L. Galvani, *Ancora sul metodo della popolazione tipo al lume della logica*, in : « Statistica, già Supplemento ai Nuovi Problemi di Politica, Storia e Economia », n. 1, Anno III, Gennaio-Marzo 1943. Quanto è affermato nel testo segue immediatamente dalle considerazioni a pag. 15 dell'estratto.

(numero maschi = numero femmine), oppure, meglio, si eseguano due confronti, sui due gruppi scelti che interessano, anziché su una sola popolazione globale che non interessa affatto. Se una vera popolazione tipo di riferimento non esiste, perchè sforzarsi di crearla? E senza una popolazione di riferimento, il problema non ha senso. Infatti, in questo caso, la probabilità indipendente:

$$_{\alpha}q = 1 - (1 - q_1)\,(1 - q_2)$$

di eliminazione per la causa α nei due gruppi non serve, perché esprime la probabilità per un individuo di essere eliminato per una almeno delle due cause, cioè, o come maschio o come femmina. Il che è assurdo, ove si tenga presente che le due eventualità, M e F, sono incompatibili, e la formula appropriata del calcolo delle probabilità fornisce:

$$_{\alpha}q = {}_{\alpha_1}p \cdot {}_{\alpha_1}q + {}_{\alpha_2}p \cdot {}_{\alpha_2}q,$$

il che è quanto dire che il calcolo delle probabilità consiglia di ricorrere al metodo della popolazione tipo. Assumendo, ad es., ${}_{\alpha_1}p = {}_{\alpha_2}p$, si torna al vecchio, e spesso ottimo metodo della media aritmetica semplice, cui avevamo poco prima ricorso.

Ma se si hanno i quozienti di mortalità per le varie categorie professionali, ad es., il caso può apparire diverso. In realtà non lo è. Se si selezionano gl'individui per professione, ogni individuo deve appartenere all'una o all'altra classe, ma la classe non è, in questo caso, un attributo proprio, non mutabile, dell'individuo, come il sesso, bensì il risultato di una scelta, più o meno libera; sempre, aprioristicamente, revocabile, e spesso, in pratica, mutata. Si potrebbe quindi non trovare assurdo il ricorso alla /2/, come indice sintetico della mortalità per professione, eliminata l'influenza della particolare composizione della popolazione per professione. Ma è facile osservare che la /2/ esprimerebbe la probabilità di essere eliminato per morte in una almeno delle s categorie professionali considerate, in s unità di tempo. Ciò porta ad assumere, più correttamente, il valore $1/s\ {}_{\alpha_1, 2, \ldots, s}q$ invece di ${}_{\alpha_1, 2, \ldots s}q$. Ed è facile ancora osservare che $1/s\ {}_{\alpha_1, 2, \ldots, s}q$ esprime la probabilità media di essere eliminato per morte in una delle s successive unità di tempo cui si riferisce la /2/, e che invece solo la prima di tali unità di tempo, interessa. Ma nella prima unità di

tempo non possono essere considerati esposti in una classe i non eliminati da un'altra classe nella stessa unità di tempo, ma tutti gl'individui componenti il gruppo iniziale degli l_o esposti al rischio α. Siamo quindi ricondotti alla media aritmetica semplice dei q_i, $M(q_i) = 1/s \, \Sigma q_i$.

Giova fare, a titolo esemplificativo, un breve esempio numerico. Una popolazione P di 2000 individui, 1000 M e 1000 F, dà in un anno 40 morti, 22 M. e 18 F. Si ha:

$$q_1 = 0{,}022, \qquad q_2 = 0{,}018,$$

da cui:

$$p_1 = 0{,}978, \qquad p_2 = 0{,}982$$

Il quoziente di mortalità generale della popolazione è:

$$q = 0{,}020.$$

La /4'/ fornisce per q i due valori estremi:

$$\min. \bar{q}\,(1{,}2) = \frac{1 - 0{,}978 \,.\, 0{,}982}{1 + 0{,}982} = \frac{0{,}039604}{1{,}982} = 0{,}01998\,;$$

$$\max. \bar{q}\,(1{,}2) = \frac{0{,}039604}{1{,}978} = 0{,}02002\,;$$

mentre la /9/ fornisce il valore intermedio

$$\bar{\bar{q}}\,(1{,}2) = \frac{0{,}039604}{1{,}980} = 0{,}0200\,..$$

La /2/ dà ${}_{1,2}q = 0{,}039604$, e corretta per l'ampiezza dell'intervallo:

$${}_{1,2}q \,/\, 2 = 0{,}019802,$$

e corretta per il numero degli esposti, dà infine la media aritmetica:

$$(q_1 + q_2) \,/\, 2 = 0{,}02\,.$$

La stessa popolazione è stata suddivisa al principio dell'anno, mediante visita medica, in due classi: *I*, ammalati 200, e *II*. sani 1800. I 40 morti dell'anno risultino ugualmente suddivisi tra queste due classi. Si ha:

$$q_I = 0{,}100, \quad q_{II} = 0{,}0(1), \quad \text{e} \quad p_I = 0{,}900, \quad p_{II} = 0{,}9\,(8).$$

Di fronte a un quoziente di mortalità generale sempre uguale a 0,02, si ha:

$$\text{min. } \bar{q}\,(\text{I, II}) = \frac{1 - 0{,}9 \,.\, 0{,}9\ (8)}{1 + 0{,}9\ (8)} = 0{,}0553\,;$$

$$\text{max. } \bar{q}\,(\text{I, II}) = \frac{0{,}11}{1{,}90} = 0{,}0579\,; \qquad \bar{\bar{q}}\,(\text{I,II}) = \frac{0{,}11}{1{,}9\ (4)} = 0{,}0566\,;$$

$${}_{1.\,II}q = 0{,}110\,; \qquad {}_{1/2\,1,\,II}q = 0{,}055\,; \qquad 1/2\,(q_I + q_{II}) = 0{,}0\ (5).$$

I valori di ${}_{1/2\,1,2}q$ in entrambi i casi non differiscono molto dagli altri, ma per il loro particolare significato non mi sembrano molto correttamente comparabili cogli altri. In entrambi i casi i valori *min.* q e *max.*, q sono molto ravvicinati rispetto ai valori q_i di partenza. Poichè nel 1° caso tali valori risultano meno disuguali che nel 2°, risulta anche, nel 1° caso, più piccola la differenza $\bar{\bar{q}} - {}_{1/2\,1,2}q$. Ma anche nel 2° essa è molta piccola. Nel 1° caso inoltre è ${}_{1/2\,1,2}q \simeq q$ mentre nel 2° il divario, è fortissimo. Ciò dipende dal fatto che nel 1° caso la popolazione tipo prescelta coincide con quella reale, mentre nel 2° ha una composizione diversissima.

Se immaginiamo che la popolazione e i morti si distribuiscano rispetto a entrambi i caratteri come è indicato nella tabella:

	A (*I*)	*S* (*II*)	Tot.
M (1)	120 (14)	880 (8)	1000 (22)
F (2)	80 (6)	920 (12)	1000 (18)
Tot.	200 (20)	1800 (20)	2000 (40)

si può eliminare l'influenza della composizione della popolazione per entrambi i caratteri, calcolando:

$$q_{1\,I} = 14 / 120 = 0{,}11\ (6)\,; \qquad p_{1\,I} = 0{,}88\ (3)$$
$$q_{1\,II} = 8 / 880 = 0{,}0\ (09)\,; \qquad p_{2\,II} = 0{,}9\ (90)$$
$$q_{2\,I} = 6 / 80 = 0{,}075\,; \qquad p_{1\,I} = 0{,}925$$
$$q_{2\,II} = 12 / 920 = 0{,}01304\,; \qquad p_{2\,II} = 0{,}98696\,;$$

e quindi, eliminando prima l'inflenza del carattere *I, II*, e poi quella del carattere 1,2 :

$$1/2\ ({}_{I,II}\,q_1) = \frac{1 - 0{,}88\ (3) \cdot 0{,}9\ (90)}{2} = \frac{1 - 0{,}8753}{2} = 0{,}06285$$

$${}_{I,II}\,p_1 = 0{,}93715$$

$$1/2\ ({}_{I,II}\,q_2) = \frac{1 - 0{,}925 \cdot 0{,}98696}{2} = \frac{1 - 0{,}9129}{2} = 0{,}04355$$

$${}_{I,II}\,p_2 = 0{,}95645$$

$$1/2\ ({}_{I,II}\,q_{1,2}) = \frac{1 - 0{,}93715 \cdot 0{,}95645}{2} = \frac{1 - 0{,}89634}{2} = 0{,}0518.$$

Eliminando invece prima l'influenza del carattere 1,2, poi quella del carattere *I, II* : »

$$1/2\ ({}_{1,2}\,q_I) = \frac{1 - 0{,}88\ (3) \cdot 0{,}925}{2} = \frac{1 - 0{,}8171}{2} = 0{,}0914$$

$${}_{1,2}\,p_I = 0{,}9086$$

$$1/2\ ({}_{1,2}\,q_{II}) = \frac{1 - 0{,}9\ (90) \cdot 0{,}98696}{2} = \frac{1 - 0{,}97799}{2} = 0{,}0110$$

$${}_{1,2}\,p_{II} = 0{,}9890$$

$$1/2\ ({}_{1,2}\,q_{I,II}) = \frac{1 - 0{,}9086 \cdot 0{,}9890}{2} = \frac{1 - 0{,}8986}{2} = 0{,}0507\ .$$

Applicando quindi successivamente la /2/ si ha un divario, tra i risultati complessivi, di 0,0011. Se ricorriamo invece alla /9/, per $s = 2$, si ha :

$${}_{I,II}\,\bar{q}_1 = \frac{1 - 0{,}8753}{1 - [0{,}88\ (3)\ 0{,}9\ (90)]/2} = 0{,}06489, \quad {}_{I,II}\,p_1 = 0{,}93611$$

$${}_{I,II}\,\bar{q}_2 = \frac{1 - 0{,}9129}{1 + [0{,}925\ 0{,}98696]/2} = 0{,}04453, \quad {}_{I,II}\,p_2 = 0{,}95547$$

$${}_{I,II}\,q_{1,2} = \frac{1 - 0{,}93511 \cdot 0{,}95547}{1{,}94529} = 0{,}05476\ ;$$

$${}_{1,2}\,\bar{q}_I = \frac{1 - 0{,}8171}{1 + [0{,}88\ (3) + 0{,}925]/2} = 0{,}09605, \quad {}_{1,2}\,p_I = 0{,}90395$$

$${}_{1,2}\,\bar{q}_{II} = \frac{1 - 0{,}9780}{1 + [0{,}9\ (90) + 0{,}98696]/2} = 0{,}01106, \quad {}_{1,2}\,p_{II} = 0{,}98894$$

$$\bar{q}_{1,2\,I,II} = \frac{1 - 0{,}90395 \,.\, 0{,}98894}{1{,}94644} \quad 0{,}05448\,;$$

e quindi un divario complessivo minore di 0,0003. Ricorrendo però alle /2/ e /9/, per $s = 4$, si ha:

$$I\,4\,(_{1,2\,I,II}\,q) = \frac{1 - 0{,}88\,(3)\cdot 0{,}9\,(9)\cdot 0{,}925\cdot 0{,}98696}{4} = \frac{1 - 0{,}7991}{4} = 0{,}0502$$

$$\bar{q}\,(_{1,2\,I,II}) = \frac{1 - 0{,}7991}{1 + 3{,}7862/4 + 5{,}3717/6 + 3{,}3846/4} = 0{,}05447$$

mentre risulta: $M\,(q_{ij}) = 0{,}05345$.

Supponiamo una seconda popolazione π, distribuita come nella tabella seguente:

	$A\,(I)$	$S\,(II)$	Tot.	Si ha:	
$M\,(1)$	300 (20)	900 (13)	1200 (33)	$q_1 = 0{,}0275$	$p_1 = 0{,}9725$
				$q_2 = 0{,}02125$	$p_2 = 0{,}97875$
$F\,(2)$	100 (10)	700 (7)	800 (17)	$q_I = 0{,}075$	$p_I = 0{,}925$
				$q_{II} = 0{,}0125$	$p_{II} = 0{,}9875$
Tot.	400 (30)	1600 (20)	2000 (50)	dai quali valori si ha:	

$I/2\,_{1,2}\,q = 0{,}02408$, $\quad \bar{q}\,(_{1,2}) = 0{,}02438$, $\quad M(q_i) = 0{,}024375$

$I/2\,_{I,II}\,q = 0{,}04328$, $\quad \bar{q}(_{I,II}) = 0{,}04425$, $\quad M(q_,) = 0{,}04375$

Considerando entrambi i caratteri, abbiamo:

$q_{1I} = 0{,}0\,(6)$, $\quad q_{2I} = 0{,}01\,(4)$, $\quad q_{1II} = 0{,}1$, $\quad q_{2II} = 0{,}01$

$p_{1II} = 0{,}9\,(3)$, $\quad p_{2I} = 0{,}98\,(5)$, $\quad p_{1II} = 0{,}9$, $\quad p_{2II} = 0{,}99$

Applicando successivamente la /2/, per $s = 2$, si ha:

$I/2\,(_{I,II}\,q_1) = 0{,}04007$; $\quad I/2\,(_{I,II}\,q_2) = 0{,}0545$;

$I/2\,(_{1,2}\,q_I) = 0{,}08$; $\quad I/2\,(_{1,2}\,q_{II}) = 0{,}01215$;

$I/2\,(_{I,II}\,q_{1,2}) = 0{,}04619$;

$I/2\,(_{1,2}\,q_{I,II}) = 0{,}04559$;

e applicando invece la /9/:

$_{I,II}\,\bar{q}_1 = 0{,}04090$; $\quad _{I,II}\,\bar{q}_2 = 0{,}05640$; $\quad _{I,II}\,\bar{q}_{1,2} = 0{,}04829$;

$_{1,2}\,\bar{q}_I = 0{,}08348$; $\quad _{1,2}\,\bar{q}_{II} = 0{,}01222$; $\quad _{1,2}\,\bar{q}_{I,II} = 0{,}04850$;

E per $s = 4$ si ha:

$$1/4\ {}_{1,2,I,II}\,q = 0{,}0451\ ; \qquad q_{(1,2,I,II)} = 0{,}04849\ ;$$

mentre infine è $M(q_{ij}) = 0{,}0477$; ed è $q = 0{,}025$.

Ricapitolando, osserviamo che il confronto dei quozienti effettivi dà, in entrambe le popolazioni: $M > F$, $A > S$, mentre tutti i metodi da noi applicati per eliminare l'influenza di un carattere, quando si considera l'altro, danno: $M > F$, $A > S$, nella 1ª popolazione, e $M < F$, $A > S$, nella 2ª.

Il confronto tra i saggi della 1ª popolazione con quelli della 2ª è eseguito nella tabella seguente. Tutti i metodi danno luogo agli stessi risultati, e ciò conforta la tesi che il miglior metodo è il più spedito, e cioè quello della media aritmetica semplice dei quozienti dei gruppi scelti. In particolare dall'esame della tabella risulta che i quozienti della popolazione P sono superiori a quelli della π (seguenti in grassetto).

Confronto tra i saggi delle due popolazioni

		effettivi			*corretti*		
		A (I)	S (II)	tot. (I,II)	$1/2\ {}_{I,II}q_i$	${}_{I,II}\bar{q}_i$	$M(q_i)$
M (1)		>	<	<	>	>	>
F (2)		<	>	>	<	<	<
tot. (1,2)		>	<	<	**>**	**>**	
$1/2\ {}_{1,2}q_j$	corretti	>	<	**>**	**>**	$s=4$	
$\bar{}_{1,2}q_j$		>	<	**>**	$s=4$	**>**	
$M(q_j)$							**>**

Se, adottando il metodo della popolazione tipo, applichiamo i quozienti della 1ª popolazione alla 2ª, e viceversa, troviamo, nel 1° caso, in cifra tonde, 61 morti, invece di 50, e, nel 2°, 38 invece di 40: anche da questo confronto risultano dunque i saggi della 1ª popolazione maggiori di quelli della 2ª.

Applicando invece i 4 quozienti delle due popolazioni alle quattro popolazioni indicate nello specchietto seguente, essi totalizzano il numero di morti indicato nelle ultime due righe. Per le popolazioni 3 o 4 i risultati sono concordati con quelli forniti dai precedenti metodi di eliminazione, e per le popolazioni 1 e 2 sono discordanti.

Quozienti		Popolazione tipo			
P	π	1	2	3	4
0,1167	0,0667	100	200	300	400
0,0091	0,0144	200	300	400	100
0,0750	0,1000	300	400	100	200
0,0130	0,0100	400	100	200	300
P		41,19	57,37	48,75	66,49
	π	43,55	58,66	37,77	51,12

Qual'è il valore di questi confronti ?

Non è dubbio che il giudizio complessivo sui quozienti di mortalità per età, o, in generale, sui quozienti di eliminazione per una causa secondo le modalità di un carattere A le quali costituiscano una successione in ciascun termine della quale possano essere eliminati per la causa α gl'individui non ancora eliminati nelle precedenti (p. es. : quozienti di eliminazione delle spose per nascita del primogenito, secondo l'età, = probabilità di nascita del primogenito, secondo l'età della madre) possa essere formulato correttamente in base alla popolazione stazionaria corrispondente alla relativa tavola di eliminazione. Quando costruire una tavola di eliminazione mettendo in fila le varie modalità del carattere A non ha senso, allora sono spesso giustificabili le formule /9/ e /10/. Queste sono certe funzioni dei ${}_{ai}q$ e corrispondono a certi scopi. Non è escluso, anzi è certo, che in relazione a altri scopi, altre funzioni saranno più appropriate. Perchè, come ho avuto già occasione di rilevare per il confronto di disuguaglianza fra n dati [1], anche la disuguaglianza fra due masse di n dati ha molteplici aspetti tutti legittimi, tra i quali vale un solo criterio di cernita : aspetti che interessano e aspetti che non interessano. I quozienti di mortalità per età di una po-

---

[1] *Recente letteratura sugl'indici di variabilità*, in : « Metron », vol. XII, n. 3, 1935.

polazione possono essere o no equivalenti (uguali o disuguali, nel loro complesso) a quelli di un'altra popolazione, a seconda che forniscono quozienti generali di mortalità uguali o diversi; o a seconda che producano numeri annui uguali o diversi di morti nelle popolazioni stazionarie da essi generate, e così via. In definitiva, è solo in base allo scopo della ricerca che esiste un criterio corretto di confronto, tutti i criteri essendo, a priori, ugualmente buoni (se servono allo scopo prefisso) o ugualmente cattivi (se non servono allo scopo prefisso). Nella quale opinione ci conforta l'autorevole parola del Gini [1].

§ 3. — I quozienti di eliminazione che abbiamo fin qui costruito per la popolazione complessiva P, secondo la causa α, rispondono allo scopo di eliminare la influenza della particolare composizione della popolazione P in classi parziali P_i omogenee rispetto ad α, cioè tali che la frequenza nell'unità di tempo del carattere A provocato da α, in ciascuna sottoclasse di P_i, sia sempre quello stesso valore ${}_{\alpha i}q$ che abbiamo assunto come misura della intensità di eliminazione della causa α in P_i. Cioè le ${}_{\alpha i}q$ debbono godere della proprietà di essere invarianti rispetto a tutte le possibili suddivisioni delle P_i in classi più piccole: se di tale proprietà non godono, i ${}_{\alpha i}q$ risultano medie ponderate di altre ${}_{\alpha i,\ldots k}q$, dipendenti dalla composizione dei gruppi P_i rispetto alle modalità dell'altro o degli altri caratteri considerati nella scomposizione nei vari P_i. Quando ciò accade, ed è la regola, le formule di cui siamo venuti discorrendo hanno un significato più ristretto, e servono a mettere in luce la intensità generale della eliminazione per la causa α, ma secondo le modalità di quel particolare carattere, che distinguono le varie classi P_i. Fin dove sia conveniente spingere il tentativo della costruzione di classi parziali omogenee rispetto alla α, è questione pratica, essenzialmente connessa alla natura della collettività P, e agli scopi della ricerca, che qui non vogliamo proporci. Occorre tener presente, in linea di massima, che, mentre la progressiva suddivisione del materiale permette la eliminazione di cause sempre più numerose di perturbazione,

---

[1] Cfr.: C. Gini, *I pericoli della Statistica*, in: «Atti della I Riunione scientifica della S. I. S., Pisa 1939, pag. 4.

quali la influenza della relativa importanza numerica delle singole classi, essa porta a determinare i rapporti elementari ${}_{\alpha i}q$ sopra numeri d'individui sempre più esigui, sì da togliere ad essi ogni carattere di stabilità, e quindi quel valore di norma che ordinariamente ci si propone con tal genere di ricerche.

A questo inconveniente della non assoluta omogeneità delle classi P_i si può anche opporre, in linea generale, la considerazione che, formate le classi P_i nelle quali più chiaramente la causa α agisce con intensità diverse, le sottoclassi nelle quali si potrebbe ancora suddividere ciascuna di queste differirebbero, a priori, molto poco l'una dall'altra nei riguardi della causa α. Se si suppone, com'è plausibile, che esse sieno anche in grande numero, è sempre possibile ammettere che i valori dei ${}_{\alpha i}q$, pur dipendendo dall'accidentale composizione delle P_i, presentino una sufficiente stabilità.

Ma un'altra ipotesi è stata anche alla base delle nostre considerazioni: che durante l'intervallo di tempo considerato la collettività P sia rimasta immutata, cioè che nessun altro individuo sia entrato a farne parte nel corso della osservazione, e nessuno di essa ne sia uscito, per causa diversa dalla α. La condizione è molto restrittiva, e difficilmente è soddisfatta in collettività reali. Soccorre però la osservazione che, ai fini della determinazione dei ${}_{\alpha_i}q$, non sono i singoli casi della collettività, che interessano, bensì il comportamento globale, nei riguardi della causa α, delle collettività parziali che la compongono. Nulla quindi risulta mutato se ad un gruppo d'individui non presentanti ancora la caratteristica A, si sostituisce, in un certo momento, un gruppo ugualmente numeroso di altri individui che presentino lo stesso comportamento rispetto ad α. Se indichiamo con $k\,(t)$ la differenza tra il numero degli entrati e il numero degli usciti per causa diversa da α, al tempo t, la collettività P_i potrà ancora dirsi chiusa, quando sussista in tutta la unità di tempo la condizione $k\,(t) = 0$ insieme alla condizione ora esposta. Rientrano generalmente in questo concetto più lato di collettività chiuse, per es. rispetto al rischio d'infortunio sul lavoro, i lavoratori delle industrie che impiegano giornalmente un numero fisso di dipendenti, trattenendone una certa aliquota costantemente a disposizione, per eventuali sostituzioni, quali, ad es., in generale, le aziende dei trasporti urbani.

Indichiamo con β il complesso delle cause diverse da α (o non α), per le quali un individuo non α può essere eliminato da P_i. Se, pur essendo valida la $k(t) = 0$ in tutta la unità di tempo, il gruppo degli eliminati per β non è , rispetto ad α, equivalente al gruppo degli entrati, allora il numero degli eliminati per α, nella unità di tempo, dalla nuova collettività P'_i, sarà diverso da quello che sarebbe stato nella collettività chiusa P_i. Per calcolare tale nuovo numero, occorrerà togliere agli $l_0\ {}_{\alpha_i}q$ eliminati nel gruppo chiuso per α, quelli che sarebbero provenuti dagli usciti per β, e aggiungere quelli che risultano eliminati dai nuovi entrati. Se indichiamo con $l_{o\ \alpha_i'}q$ gli eliminati da l_0 per α fino al momento t, risulteranno $l_0\,({}_{\alpha_i}q - {}_{\alpha_i'}q)$ gli eliminati nel gruppo chiuso dall'istante t in poi, $l_0\,(1 - {}_{\alpha_i'}q)$ gli esposti all'inizio dell'intervallo $1 - t$, e quindi ${}_{\alpha_i''}q = ({}_{\alpha_i}q - {}_{\alpha_i'}q) / (1 - {}_{\alpha_i'}q)$ il quoziente di eliminazione nella collettività chiusa dall'istante t alla fine del periodo di osservazione. Se indichiamo con θ la frazione dei sopravviventi (non α), sostituiti nell'istante t per effetto del movimento che avviene nel gruppo, e con ${}_{\alpha\beta}q$ il quoziente di eliminazione relativo ai nuovi entrati, per tutto il tempo della loro osservazione, sarà $\theta l_0\,(1 - {}_{\alpha_i'}q)\ {}_{\alpha_i''}q$ il numero dei non più eliminati dal gruppo chiuso e $\theta l_0\,(1 - {}_{\alpha_i'}q)\ {}_{\alpha\beta}q$ il numero degli eliminati provenienti dai nuovi entrati. In totale, nella unità di tempo, il numero degli eliminati nel gruppo aperto sarà :

$$l_{o\ \alpha_i}q - \theta l_0\,(1 - {}_{\alpha_i'}q)\ {}_{\alpha_i''}q + \theta l_0\,(1 - {}_{\alpha_i'}q)\ {}_{\alpha\beta}q =$$

$$= l_{o\ \alpha_i'}q + (1 - {}_{\alpha_i'}q)\,\theta l_0\,({}_{\alpha\beta}q - {}_{\alpha_i''}q),$$

e quindi il relativo quoziente di eliminazione :

$${}_{\alpha_i}q + \theta\,(1 - {}_{\alpha_i'}q)\ ({}_{\alpha\beta}q - {}_{\alpha_i''}q)\,.$$

Come si vede, tale quoziente, a parità di ${}_{\alpha_i}q$, di θ e di ${}_{\alpha\beta}q$, dipende dalla distribuzione degli eliminati del gruppo chiuso nei due intervalli nei quali l'istante t divide la unità di tempo. Se movimento avviene ad ogni istante, è tutta la legge temporale di

eliminazione che contribuisce alla formazione del quoziente di eliminazione della nuova collettività, e non solo quella del gruppo iniziale, ma anche di tutti i successivi gruppi di entrati, se anche questi sono soggetti a eliminazioni per la causa β, dopo il loro ingresso nella collettività P_i.

La legge temporale di eliminazione da un gruppo è definita dai quozienti di eliminazione relativi ai successivi intervalli di tempo Δt. Se opera la sola causa α di eliminazione, indicato con $l(t)$ i sopravviventi (non eliminati) del gruppo iniziale l_0 al tempo t, e con $\mu(t)\,\Delta t$ il quoziente di eliminzione per l'intervallino Δt, sarà $\frac{l(t) - l(t+\Delta t)}{l(t)} = \mu(t)\,\Delta t$, e quindi, se esiste il limite del rapporto $\frac{l(t+\Delta t) - l(t)}{\Delta t}$ per $\Delta t \to 0$, risulterà

$$\mu(t) = -\,dlog\,.\,l(t)$$

e quindi: $l(t) = l(t_1)\; e^{-\int_{t_1}^{t} {}_{\alpha}\mu(u)\,du}$.

Se opera anche la causa di eliminazione β (non α), in ogni intervallino Δt gli eliminati per le due cause saranno:

$$l(t) - l(t+\Delta t) = l(t)\,[\,{}_{\alpha}\mu(t) + {}_{\beta}\mu(t)]\,\Delta t,$$

dalla quale si ricava, nella ipotesi precedente:

$$l(t) = l(t_1)\; e^{-\int_{t_1}^{t} [\,{}_{\alpha}\mu(u) + {}_{\beta}\mu(u)\,]\,du}\ (1). \qquad /11/$$

Indichiamo con $I(u)$ i nuovi entrati in P all'istante u. I relativi quozienti istantanei di eliminazione (per α e β), al momento v, saranno funzioni e dell'istante v e dell'istante u dell'ingresso del gruppo: li indicheremo pertanto con $\mu(u, v)$.

I sopravviventi all'istante t saranno:

$$E(u, t) = I(u)\; e^{-\int_{u}^{t} [\,{}_{\alpha}\mu(u,v) + {}_{\beta}\mu(u,v)\,]\,dv}$$

e di questi, gli eliminati per la causa α al tempo t saranno: $E(u, t)\,{}_{\alpha}\mu(u, t)\,dt$, sicchè tutti gli eliminati durante il tempo $t, t+\Delta t$ risultano:

$${}_{\alpha}m\,(t)\,dt = \int_{t_1}^{t} E(u, t)\,{}_{\alpha}\mu(u, t)\,du\,.\,dt\ .$$

---

(1) F. P. Cantelli, *Genesi e costruzione ecc.*, già cit.

Il rapporto $\frac{m\,(t)\,dt}{\int_{t_1}^{t} E\,(u,\,t)\,du}$ del numero degli eliminati per α nello intervallo $t, t + dt$ al numero degli esposti al tempo t definisce la forza istantanea di eliminazione per la causa α, nella collettività risultante dalle leggi indicate degl'ingressi e delle eliminazioni. Indicata questa con ${}_{\alpha}\mu(t)$, si ha pertanto:

$$ {}_{\alpha}\mu\,(t) = \frac{\int_{t_1}^{t} {}_{\alpha}\mu\,(u,\,t)\,E\,(u,\,t)\,du}{E\,(u,\,t)\,du} \qquad /12/ $$

La /12/ vale anche per un gruppo comunque aperto (cioè non vincolato alla condizione $k\,(t) = O$), perchè non abbiamo fatto ricorso a questa condizione per determinarla. Possiamo dire pertanto che, in un gruppo aperto, il tasso istantaneo di eliminazione per una causa α dipende, e dai tassi istantanei relativi ai singoli gruppi di entrati, e dall'ammontare di tali gruppi. Posto $E\,(t) = \int_{t_1}^{t} E\,(u,\,t)\,du$, per il numero degli eliminati dal gruppo, nell'intervallo t_1, t_2, risulta l'espressione:

$$ \int_{t_1}^{t_2} m\,(t)\,dt = \int_{t_1}^{t_2} E\,(t)\;{}_{\alpha}\mu\,(t)\,dt \qquad /13/ $$

Se $k\,(t) = O$, per il che si richiede che sia per ogni t:

$I(t) = \int_{t_1}^{t} E(u,t)\;{}_{\beta}\mu\,(u,t)\,du$, risulta $E(t) = l(t) = l(t_1)\;e^{-\int_{t_1}^{t} {}_{\alpha}\mu(u)\,du}$ /14/

e la collettività aperta P equivale a un gruppo chiuso nel quale la legge di eliminazione sia definita dalla /12/: risulta pertanto giustificata la definizione

$$ {}_{\alpha}q(t_1,\,t_2) = m(t_1,t_2)\;/\;l\,(t_1) \qquad /15/ $$

per il quoziente di eliminazione nell'intervallo $t_1\,t_2$, e

$$ {}_{\alpha}q = m(t_1,t_2)/(t_2 - t_1)\;l\,(t_1) \qquad /16/ $$

per il quoziente medio di eliminazione nella unità di tempo, relativo alla collettività concreta P. Queste formule hanno anche il significato di probabilità indipendenti, nel senso sopra definito. Ad esse resta collegato il tasso centrale di eliminazione:

$$ {}_{\alpha}q = m(t_1,t_2)\,/\int_{t_1}^{t_2} l(t)\,dt \qquad /17/ $$

esprimente il quoziente generale di eliminazione nella popolazione stazionaria generata dalla funzione ${}_{\alpha}\mu(t)$. E poichè nel nostro caso è, per qualunque t, $l(t) = E(t)$, $l(t)$ è funzione statisticamente rilevabile.

Ma se la funzione $k(t)$ non è identicamente nulla in tutto l'intervallo, come si dovrà definire il quoziente di eliminazione per la causa α nella collettività P? È questo un problema che ha avuto una soluzione classica. Vediamo anzitutto quale è stata questa soluzione.

Ad ogni istante t è come se entrassero nella collettività P un numero $k(t)\,dt$ individui, uguale alla differenza tra il numero degli entrati e quello degli usciti per causa $\beta \neq \alpha$. Indicato con ${}_{\alpha}q(t,t_2)$ la probabilità di eliminazione da questo gruppo nell'intervallo t, t_2, per la causa α, risulta $k(t)\,{}_{\alpha}q(t, t_2)\,dt$ il numero degli eliminati di P dai $k(t)\,dt$ aggiunti nell'intervallino $t, t + dt$. Tutti gli eliminati di P per la causa α saranno quindi

$$\int_{t_1}^{t_2} m(t)\,dt = E(t_1)\,{}_{\alpha}q(t_1t_2) + \int_{t_1}^{t_2} k(t)\,{}_{\alpha}q(t,t_2)\,dt \qquad /18/$$

Posto:

$${}_{\alpha}q(t,t_2) \,/\, {}_{\alpha}q(t_1,t_2) = f(t),$$

risulta:

$$\int_{t_1}^{t_2} m(t)\,dt = q(t_1,t_2)\left[E(t_1) + \int_{t_1}^{t_2} k(t)\,f(t)\,dt\right],$$

dalla quale:

$${}_{\alpha}q(t_1,t_2) = \frac{\int_{t_1}^{t_2} m(t)\,dt}{E(t_1) + \int_{t_1}^{t_2} k(t)\,f(t)\,dt} \qquad /19/$$

Ad ogni determinazione della funzione $f(t)$ corrisponde una formula per ${}_{\alpha}q(t_1,t_2)$.

Cerchiamo d'inquadrare la equazione /18/ nella nostra trattazione.

Se indichiamo con $U(t)$ gli usciti al tempo t per causa $\neq \alpha$, è $k(t) = I(t) - U(t)$. Gli $U(t)$ provengono da tutti gli entrati nell'intervallo t_1t, secondo le varie leggi di eliminazione rappresentate dalla funzione ${}_{\beta}\mu(u,t)$. Degli $I(u)\,du$ entrati nell'intervallino $u, u + du$, escano al tempo t numero $U(u, t)\,dudt$ individui, i quali avrebbero contribuito con $U(u, t)\,q(u, tt_2)\,dudt$ individui

al numero complessivo degli eliminati dalla P per la causa α. Il numero pertanto dei non eliminati dalla P a causa degli usciti al tempo t per causa non α (provenienti da tutti gli entrati fino al tempo t), è $dt \int_{t_1}^{t} U(u,t)\ {}_{\alpha}q(u, tt_2)\ du$, e il numero complessivo dei non eliminati durante tutto l'intervallo:

$\int_{t_1}^{t_2}\left[\int_{t_1}^{t} U(u, t)\ {}_{\alpha}q\ (u, tt_2)\ du\right] dt$, mentre $U(t) = \int_{t_1}^{t} U\ (u, t)\ du$ è il numero degli usciti al tempo t. Sarà quindi:

$$\int_{t_1}^{t_2} m(t)\ dt = E(t_1)\ {}_{\alpha}q(t_1 t_2) + \int_{t_1}^{t_2}\left[I(t)\ {}_{\alpha}q(t, t_2) - \int_{t_1}^{t} U(u, t)\ {}_{\alpha}q\ (u, tt_2)\ du\right] dt \qquad /20/$$

e questa equazione risul a uguale alla /18/ solo se è soddisfatta, per ogni t, la relazione

$$\int_{t_1}^{t} U\ (u, t)\ \ q\ (u, tt_2)\ du = {}_{\alpha}q\ (t, t_2)\ U(t) \qquad /21/$$

Questa equazione non può essere identicamente soddisfatta se non è ${}_{\alpha}q\ (u, tt_2) = {}_{\alpha}q\ (t, t_2)$ per ogni u. Ciò significa che i tassi istantanei di eliminazione non dipendono dalla variabile u, e quindi che tutte le varie schiere di entrati presentano la stessa legge di eliminazione. Cioè la P costituisce una collettività selezionata.

La /19/ è pertanto legittima solo per la costruzione di tavole selezionate. La ipotesi $\frac{\partial\ {}_{\alpha}\mu\ (u, t)}{\partial\ u} = O$ /22/, che caratterizza la omogeneità dei vari gruppi, di entrati, risulta implicita nella dimostrazione che della /18/ dà il Cantelli [1], ma passa ordinariamente inosservata [2].

---

[1] F. P. Cantelli, *Genesi e costruzione ecc.*, già cit., pagg. 12, 28 e seguenti.

[2] Vedi: A. Del Chiaro, *Sulle tavole di mortalità*, in: « Giornale dell'Istituto Italiano degli Attuari », Anno XII, n. 2, Aprile 1941; e: *Sui tassi centrali di mortalità*, ibid., Anno XII, nn. 3-4, Luglio-Ottobre 1941. Così anche il Wolfenden, nel suo, sotto alcuni aspetti pur notevole lavoro: *On the formulae for calculating the exposed to risk in constructing mortality and other tables from the individual records of insured lives*, in:

Quando è legittima, la (19) rappresenta la probabilità indipendente di eliminazione da P per α, e questa è indipendente dall'ammontare assoluto degli entrati e degli usciti per causa $\neq \alpha$, sicchè ci si può , almeno in linea teorica, sempre ricondurre al caso $k(t) = 0$, nel quale valgono le /15/ e /16/. Al numeratore dei quozienti a 2° membro di queste formule non va più posto, naturalmente, il numero ${}_{\alpha}m(t_1\, t_2)$ degli eliminati effettivamente da P durante l'intervallo, ma quello virtuale degli eliminati nella ipotesi $I(t) = U(t)$. Il calcolo di questo numero si può effettuare diversamente a secondo degli elementi statistici a disposizione. Se si conosce la distribuzione degli eliminati e degli esposti, è applicabile la /2/, in una delle sue varie forme, assumendo, come ${}_{\alpha}q_i$ i rapporti ${}_{i}m(\Delta t)/{}_{i}E(\Delta t)$ degli eliminati nei successivi intervallini parziali nei quali può dividersi l'intervallo $t_1\, t_2$, al numero degli esposti all'inizio degl'intervallini stessi; oppure i rapporti ${}_{i}m(t_1, \Delta t) / {}_{i}E(t_1, \Delta t)$ degli eliminati provenienti dal gruppo degli esistenti all'inizio t_1 dell'intervallo $t_1\, t_2$ agli esposti all'inizio dell'intervallino, provenienti da detti $l(t_1)$, cioè i sopravviventi, non solo alla causa di eliminazione α, ma alla causa totale di eliminazione $\alpha + \beta$. Le due determinazione di ${}_{\alpha}q_i$, identiche in teoria, debbono essere non sistematicamente diverse in pratica, se deve ritenersi valida la ipotesi che la P costituisca una collettività selezionata.

Per porre la /2/ sotto la forma /5/, corrispondente alla /15/, occorre procedere al calcolo del numero virtuale degli eliminati per α dal gruppo iniziale, nell'assenza di ogni altra causa β, di eliminazione, in ogni intervallino Δt, e questo si otterrà moltiplicando il numero degli eliminati per α, per il rapporto tra i sopravviventi nella legge pura di eliminazione e il numero degli effettivi esposti al rischio; il numero dei sopravviventi all'inizio

---

« Transactions Actuarial Society of America », vol. XLIII, n. 108. E vedi anche l'OTTAVIANI: *Sulle tavole di mortalità*, in: « Giornale dell'Istituto Italiano degli Attuari », Anno XIII, n. 1-4, Gennaio-Ottobre 1942, che scrive: « ...le considerazioni (che muovono dalla formula /19/) si estendono alle tavole selezionate ed aggregate di eliminazione, le quali forniscono la probabilità di eliminazione per una causa determinata, e alle tavole di mutualità, secondo quanto ha già fatto il Cantelli nella sua memoria del 1914 ». Ma non pare che il Cantelli abbia esteso, nella Memoria del '14, le proprie considerazioni alle tavole non selezionate.

dell'intervallino i.esimo nella legge pura di eliminazione risultando dal numero degli esposti al rischio all'inizio dell'intervallo $t_1\ t_2$ diminuito del numero virtuale degli eliminati in tutti gl'intervallini precedenti.

Il numero virtuale degli eliminati nell'intervallino i. esimo risulta dunque:

$$ {}_{i}m'\ (\Delta\ t) = {}_{i}m\ (\Delta\ t)\ {}_{i}l\ (t) / {}_{i}E\ (t) \qquad /22/ $$

dove

$$ {}_{i}l\ (t) = l\ (t_1) - \sum_{1}^{i-1} {}_{j}m'\ (\Delta\ t) \qquad /23/ $$

Se si conosce la distribuzione degli eliminati provenienti dal gruppo iniziale degli esposti, si ha:

$$ {}_{i}m'\ (t_1, \Delta\ t) = {}_{i}m\ (t_1, \Delta\ t)\ {}_{i}l\ (t_1, t) / {}_{i}E(t_1, t) \qquad /22'/ $$

dove

$$ {}_{i}l\ (t_1,t) = l\ (t_1,t) - \sum_{1}^{i-1} {}_{j}m'\ (t_1, \Delta\ t) \qquad /23'/ $$

Le /22/ e /23/, /22'/ e/23'/ debbono condurre a numeri complessivi di eliminati praticamente uguali:

$$ \Sigma\ {}_{i}m'\ (\Delta\ t) \simeq \Sigma\ {}_{i}m'\ (t_1, \Delta\ t), $$

e da questa ultima relazione segue che, posto

$$ m'\ (t_1\ t_2) = \Sigma_i\ m'\ (\Delta\ t), $$

$$ m'\ (t_1, t_1\ t_2) = \Sigma_i m'\ (t_1, \Delta\ t), $$

è:

$$ {}_{\alpha}q\ (t_1\ t_2) = m'\ (t_1\ t_2) / l\ (t_1) \qquad /24/, $$

$$ {}_{\alpha}q\ (t_1\ t_2) = m'\ (t_1, t_1\ t_2) / l\ (t_1) \qquad /24'/ $$

Qualora il movimento che avviene nella collettività sia di scarsa importanza, relativamente alla massa degli esistenti all'inizio del periodo, e i ${}_{\alpha}q_i$ non troppo forti, talché trascurando gli eliminati per α provenienti dall'eccesso $k\ (t)$ degli entrati sugli usciti non si commetta errore troppo sensibile, può essere lecito sostituire ai valori ${}_{i}m'\ (\Delta\ t)$ i valori approssimari:

$$ {}_{i}m''\ (\Delta t) = {}_{i}m\ (\Delta\ t) \frac{{}_{i}E\ (\Delta\ t) - \sum_{1}^{i-1} {}_{j}k\ (\Delta\ t)}{{}_{i}E\ (\Delta\ t)} \qquad /25/, $$

e in base a questa, totalizzare il numero degli eliminati:

$$m'' (t_1 t_2) = \Sigma \, {}_{\Delta}m'' (\Delta t)$$

che fornisce il valore approssimato

$$ {}_{a}q'' (t_1 t_2) = m'' (t_1 t_2) \, / \, l (t_1). \qquad /26/$$

Ove gli elementi statistici a disposizione non permettano l'uso di queste formule, si può ricorrere ai metodi classici delle durate esatte, medie, ecc.

A titolo di esempio, calcoliamo la probabilità del concepimento del primogenito durante il 1° anno di matrimonio, e la probabilità di un concepimento durante l'8° anno di matrimonio, per l'insieme delle 90 donne sposate nel 21° anno di età, comprese nel gruppo delle famiglie sovrane di Europa, mediatizzate tedesche e principesche italiane, francesi, inglesi, russe, austriache, ungheresi e polacche, riportate nelle annate 1891-927 dell'Almanacco di Gotha. Il materiale, messo gentilmente a disposizione dal prof. Savorgnan, è quello stesso, già usufruito dai proff. Savorgnan e Gini per una serie di articoli che elenco in nota [1], comprendente la discendenza fino al 31/12/26 dei matrimoni contratti nel periodo 1890-909 più, per le case mediatizzate tedesche, nuovi dati, anch'essi raccolti e gentilmente forniti dal prof. Savorgnan, riguardanti i figli nati fino al 16/9/40 da matrimoni contratti negli anni 1910-21.

---

[1] F. Savorgnan, *Nuzialità e fecondità delle Case Sovrane d'Europa*, in: « Metron », vol. III, n. 2; *La fecondità delle aristocrazie*, I, ibid., vol. III, n. 3/4; II, ibid., vol. IV, n. 3/4; III, ibid., vol. V, n. 1; IV, ibid., vol. IX, n. 1. C. Gini, *Su la curva della fecondità matrimoniale secondo l'età*, XXII Riunione dell'Istituto Internazionale di Statistica, Londra, 1934; *Su la determinazione dei quozienti di eliminazione, e in particolare sui metodi delle durate esatte e delle durate medie, nell'ipotesi di saggi istantanei di eliminazione costanti*, in: « Metron », vol. XII, n. 3. 1935.

Tab. 1.

(1) Durata del matrimonio	(2) $\sum_{t_1}^{t} U(\Delta t)$	(3) $E(t)$	(4) Concepimenti
0	—	90	48
0,54	1	41	
0,56	2	40	2
1	2	38	
7	6	84	
7,08	7	83	10
7,15	8	82	
7,66	9	71	4
8	9	67	

Nella 3ª colonna è indicato il numero dei matrimoni con durata maggiore di quella a fianco indicata. Come epoca della eliminazione della donna per scioglimento del matrimonio è stata assunta, seguendo l'esempio del Gini, la data della morte del marito e del divorzio o separazione, e quella corrispondente a 9 mesi prima la data della morte della donna o della fine del periodo di osservazione, non essendo stati considerati che i concepimenti seguiti da nascita.

Nella 4ª colonna è indicato il numero dei concepimenti avvenuti nell'intervallo tra due eliminazioni per scioglimento del matrimonio. Durante il 1° anno di matrimonio, due donne hanno concepito due volte, ma noi abbiamo considerato solo il concepimento del primogenito. Nell'8° anno di durata, nessuna donna presenta due concepimenti.

Poichè tutte le donne sono state osservate dalla data del matrimonio, non esistono ingressi durante i singoli anni di durata: le formule /22/ e /23/ coincidono con le /22'/ e /23'/, e nella /25/, invece di sottrarre $\sum_{i}^{t} k(\Delta t)$, occorre aggiungere $\sum_{i}^{t} U(\Delta t)$.

Per la probabilità del concepimento del primogenito, in base alla /24/, si ha:

$$q(20,0 - 1) = \frac{48 + 2 \cdot 42 / 40}{90} = \frac{50,10}{90} = 0,55\ (6),$$

e in base alla /2'/:

$$q(20,0 - 1) = 1 - \frac{42}{90} \cdot \frac{38}{40} = \frac{50,10}{90} = 0,55\ (6),$$

mentre i metodi delle durante esatte e medie forniscono entrambi uno stesso risultato leggermente diverso:

$$q\,(20{,}0 - 1) = \frac{50}{90 - 0{,}90} = 0{,}5612\ .$$

Per la probabilità di un concepimento (di qualsiasi ordine di generazione) nell'8° anno di matrimonio, si ha:

$$q\,(20{,}7 - 8) = \frac{10\ .\ 84\ /82 + 4(84 - 10\ .\ 84/82)\ /\ 71}{84} =$$

$$\frac{10 + 4\ (82 - 10)\ /71}{82} = 0{,}1714,$$

in base alla /24/, e approssimando, in base alla /25/:

$$q\,(20{,}7 - 8) = \frac{10\ .\ 84\ /\ 82 + 4\ (84 - 10)\ /71}{84} =$$

$$\frac{10 + 4\ (82\ /84)\ (74\ /\ 71)}{82} = 0{,}1716,$$

mentre la /2'/ dà:

$$q\,(20{,}7 - 8) = 1 - (72/82)\ (\ 67/71) = 0{,}1714\ ;$$

il metodo delle durate esatte:

$$q\,(20{,}7 - 8) = \frac{14}{84 - 0{,}92 - 0{,}85 - 0{,}34} = 0{,}1710,$$

e il metodo delle durate medie:

$$q\,(20{,}7 - 8) = \frac{14}{84 - 0{,}95 - 0{,}85 - 0{,}35} = 0{,}1710.$$

La formula /25/ dà risultati meno discosti dalle esatte /2'/ e /24/ che non i metodi delle durate esatte e medie, che, nel 1° caso, risultano approssimati per eccesso, nel 2°, più decisamente, per difetto [1].

---

[1] Non applicabile al nostro materiale risulta la formula /13/ dell'Ottaviani, (*Sulle tavole di mortalità*, già cit.), che pure ha lo stesso ristretto campo di applicazione delle /2'/ e /24/, perchè esige la conoscenza della distribuzione degli eliminati lungo tutto l'intervallo. Ma anche nei casi meno semplici del nostro, quando ingressi ed egressi si succedono quasi continuamente nel gruppo, non si vede la ragione per la quale si dovrebbe giungere alla determinazione di q attraverso il calcolo di $\int \mu\,(u)\,du$, quando q si può calcolare direttamente.

Fuori dalla ipotesi, finora considerata, di una collettività selezionata, non esiste, nella collettività, una unica legge temporale di eliminazione preesistente alla formazione della P, ma una legge temporale, definita dalla /12/, si forma proprio in virtù della dinamica degl'ingressi e degli egressi. Questa legge risultante varia non soltanto se varia qualcuna delle leggi componenti, ma anche se varia l'importanza numerica dei singoli gruppi d'entrati o d'usciti.

Non può pertanto esistere una probabilità di elimina indipendente dalla particolare composizione della P reale, istante per istante. Nulla vieta però di definire la

$$q\,(t_1, t_2) = 1 - e^{-\int_{t_1}^{t_2} m\,(t)\,dt / E\,(t)} \qquad /27/$$

quale probabilità indipendente di eliminazione, nel senso che essa dà la probabilità di eliminazione in un gruppo chiuso nel quale operi la legge reale di eliminazione della P. E vale allora anche la /19/, derivata dalla /17'/, quando si sostituisca al numeratore non il numero degli eliminati dalla reale P, ma il numero degli eliminati dalla popolazione stazionaria generata dalla legge definita dalla /12/. Ma la collettività P non può che considerarsi somma di tutte le collettività $P\,(t)$, composte ciascuna di $E\,(t)$ individui, dei quali $m\,(t)\,dt = {}_{\alpha}\mu\,(t)\,E\,(t)\,dt$ risultano eliminati per α nell'intervallo $t,\ t + dt$. Ad essa si può associare pertanto la probabilità media

$${}_{\alpha}s = \frac{\int_{t_1}^{t_2} {}_{\alpha}\mu\,(t)\,E\,(t)\,dt,}{\int_{t_1}^{t_2} E\,(t)\,dt} \qquad \text{cioè} \qquad {}_{\alpha}s = \frac{m\,(t_1, t_2)}{\int_{t_1}^{t_2} E\,(t)\,dt} \qquad /28/$$

o saggio medio di eliminazione, che corrisponde anche alla aspettativa della frequenza di eliminazione per α, quando si eseguono $E\,(t)$ osservazioni, in blocco oppure no, in ciascuna delle collettività $P\,(t)$, ove si suppongano i singoli casi ugualmente osservabili (schema di Poisson) (1).

---

(1) Nulla osta, in teoria, di ricorrere anche in questo caso, al tasso centrale di eliminazione (formula /17/), ma poichè gli $l\,(t)$ non sono più statisticamente rilevabili, in quanto non coincidono cogli $E\,(t)$, la formula /17/ risulta di nessuna utilità.

§ 4. — La /28/ è tutt'altro che nuova. La trovo, per esempio, in un lavoro dello Czuber del 1910 (1), ed è stata recentemente rimessa in onore dal Gini (2), ma nessun Autore, ch'io sappia, aveva messo in luce come, per il suo significato di probabilità media, sia formula legittima, anzi la sola formula legittima per la determinazione della probabilità media di eliminazione in collettività comunque aperte.

Occorre però fare una osservazione: la s_t è una media pondo dei ${}_a\mu(t)$: i ${}_a\mu(t)$ sono probabilità relative a intervallini dt di tempo, ragguagliate all'ampiezza dt degl'intervallini. Se tali probabilità sono $> dt$ risulterà ${}_a\mu(t) > 1$. e se parecchi dei ${}_a\mu(t) > 1$, può anche risultare $s_t > 1$. Valori di $s_t > 1$ dipendono dal fatto che, come ha pur rilevato il Gini (2), i ${}_a\mu(t)$ sono probabilità istantanee, ma riferite alla unità di tempo, e non è strano che una probabilità di 1/100 in un giorno, diventi 3,65 in un anno. Ed in effetti, in una collettività aperta nella quale il numero degli esposti si mantenga rigorosamente costante, con sostituzione degli eliminati per la causa in esame e per le eventuali altre cause, per ogni 100 esposti al rischio si avranno effettivamente 365 eliminati in un anno. Gli è che la difficoltà che s'incontra nel definire il concetto di probabilità in un insieme aperto, è difficoltà reale (3), e non sparisce colla introduzione della finzione della probabilità istantanea. Una disanima approfondita del concetto non è nei fini di questo studio. Dal punto di vista pratico, qui basti osservare che l'ampiezza dell'intervallino che definisce i ${}_a\mu(t)$ non

---

(1) E. Czuber, *Wahrscheinlichkeitsrechnung und ihre Anwendung auf Fehlerausgleichung Statistik und Lebensversicherung*, II Bd, Teubner, Lipsia 1910, pag. 85, formula 13.

(2) C. Gini, *Sulla determinazione dei quozienti di eliminazione ecc.*, già cit.

(3) È per questa ragione che non posso concordare col Wolfenden, (*On the formulae*, ecc., già cit.) nella opinione che la costruzione delle tavole di mortalità « requires only the understanding of a few quite simple concepts ». Molto più spesso di quanto si possa immaginare accade che i concetti che sembrano più chiari sono i più oscuri e causa di tutte le difficoltà. Può darsi che sforzarsi di dare una definizione rigorosa del concetto di esposti al rischio provochi perplessità e confusione negli studenti, ma proprio ciò dovrebbe indurre gl'insegnanti ad una maggiore riflessione sul concetto, perchè è inammissibile che un concetto preciso non possa essere reso chiaramente.

scende mai al disotto di un minimo Δt, e che la misura di tale ampiezza aumenta riducendo l'unità di misura. Basterà quindi ridurre opportunamente l'unità di tempo per rendere tutti i ${}_{\alpha}\mu\,(t) < 1$, e quindi anche s_t, che è media interna dei ${}_{\alpha}\mu\,(t)$.

Il calcolo della /28/ non richiede nessuna ipotesi, e si esegue mettendo al denominatore la somma delle durate di esposizione al rischio dei singoli componenti la collettività, cioè:

$$ {}_{\alpha}s = \frac{m\,(t_1,\,t_2)}{(t_2 - t_1)\,l\,(t_1) + \int_{t_1}^{t_2}\left[k\,(t) - m\,(t)\right](t_2 - t)\,dt} \qquad /28'/ $$

La /28'/ è la formula delle durate esatte, del Gini.

Poniamo ora:

$$ \frac{\int_{t_1}^{t_2} {}_{\alpha}\mu\,(t)\,E\,(t)\,dt}{\int_{t_1}^{t_2} {}_{\alpha}\mu\,(t)\,dt} = {}_{\alpha}\bar{E} \qquad /29/, \quad \text{e} \quad \frac{\int_{t_1}^{t_2} E\,(t)\,dt}{\int_{t_1}^{t_2} dt} = \bar{E}. \qquad /30/ $$

Risulta:

$$ \int_{t_1}^{t_2} {}_{\alpha}\mu\,(t)\,dt = \frac{m\,(t_1\,,\,t_2)}{{}_{\alpha}\bar{E}\,(t_1\,,\,t_2)} = \frac{m\,(t_1\,,\,t_2)}{(t_2 - t_1)\,\bar{E}\,(t_1\,,\,t_2)} \cdot $$

$$ \cdot \frac{\bar{E}\,(t_1\,,\,t_2)}{{}_{\alpha}\bar{E}\,(t_1\,,\,t_2)}\,(t_2 - t_1) = (t_2 - t_1)\,\frac{\bar{E}}{{}_{\alpha}\bar{E}}\,{}_{\alpha}s, $$

da cui:

$$ {}_{\alpha}q\,(t_1\,,\,t_2) = 1 - e^{-(t_2 - t_1)\,{}_{\alpha}s\,\frac{\bar{E}}{{}_{\alpha}\bar{E}}}. \qquad /31/ $$

Questa formula esprime la relazione che passa tra la probabilità indipendente di eliminazione, definita dalla /27/, e la probabilità media /28/, sotto una forma che mette in luce la seguente proprietà: a parità di probabilità media la probabilità indipendente di eliminazione è funzione del rapporto $\bar{E} \,/\, {}_{\alpha}\bar{E}$ della media semplice degli esposti alla media ponderata degli stessi, con pesi ${}_{\alpha}\mu\,(t)$; per una data legge di eliminazione è massima quando i valori di $\bar{E}\,(t)$ sono contrograduati con i corrispondenti ${}_{\alpha}\mu\,(t)$, minima quando sono invece cograduati.

Sia nella ipotesi ${}_{\alpha}\mu\,(t) = cost.$, che nella ipotesi $E\,(t) = cost.$, risulta $\bar{E} = {}_{\alpha}\bar{E}$, e quindi la /32/ assume la forma

$$ {}_{\alpha}q = 1 - e^{-(t_2 - t_1)\,{}_{\alpha}s} \qquad /32/ $$

Nella 1ª delle due ipotesi risulta anche

$$ {}_{a}s = {}_{a}\mu, \qquad /33/ $$

e, in entrambe:

$$ \int_{t_1}^{t_2} {}_{a}\mu\,(t)\,dt = (t_2 - t_1)\,{}_{a}s\,. \qquad /34/ $$

Fuori di queste due ipotesi, la /32/, che potremo scrivere anche, più brevemente, ${}_{a}q = 1 - e^{-{}_{a}s}$ /35/, intendendo ${}_{a}q$ e ${}_{a}s$ relativi alla stessa unità di tempo, costituisce una comoda relazione approssimata per la determinazione di ${}_{a}q$, particolarmente utile quando le distribuzioni degli $m\,(t)$ e degli $E\,(t)$ fanno ritenere che il rapporto $\bar{E} \,/\, {}_{a}\bar{E}$ debba mantenersi sufficientemente vicino a 1, evitando di ricorrere alla esatta, ma scomoda in pratica, /27/, se non, eventualmente, come controllo dei risultati forniti dalla /35/ quando $\bar{E} \,/\, {}_{a}\bar{E} \simeq 1$ [1].

La formula /35/ è stata determinata, per ${}_{a}\mu = cost.$ (nel quale caso è esatta) dal Gini [2] con un procedimento quanto mai semplice ed elegante. Questo Autore ha dedotto la espressione di ${}_{a}q$ da quella di ${}_{a}s$, (che chiama saggio medio di eliminazione) per un gruppo chiuso, ed ha osservato che la formula vale anche nel caso di un gruppo aperto.

Il solo, che io sappia, che si è occupato prima del Gini, della ipotesi del tasso istantaneo costante, è stato il Wittstein, il quale però, procedendo per la via classica, prima determina la probabilità di eliminazione in un intervallo $(x,1)$ contenuto nell'intervallo unitario $(0,1)$, e poi calcola separatamente i sopravviventi

---

[1] È vero senza dubbio che il saggio medio di eliminazione espresso da $s = \dfrac{\int E\,(t)\,\mu\,(t)\,dt}{\int E\,(t)\,dt}$ non può ritenersi, in generale, uguale al tasso centrale di eliminazione, espresso da $q = \dfrac{\int l\,(t)\,\mu\,(t)\,dt}{\int l\,(t)\,dt}$, e non so se qualcuno abbia mai asserito il contrario. Certo la confusione è anche possibile quando si adoperano gli stessi simboli per le quantità che compaiono nelle rispettive formule, e per questo motivo mi sembra che ragionevolmente insista il Del Chiaro sull'argomento (*Sui tassi centrali di mortalità*, già cit., pag. 219). Ma per evitare che, confondendo i due, alcuni attribuiscano al 1° gl'inconvenienti del 2°, che, dal punto di vista pratico, non sono pochi.

[2] C. Gini, su *La determinazione dei quozienti*, ecc., già cit.

degli esposti all'inizio dell'intervallo, e degli entrati e degli usciti, nella ipotesi $I(t) = cost.$ e $U(t) = cost.$ Poichè per

$$_{\alpha}\mu\,(t) = {}_{\alpha}\mu, \qquad \text{dalla} \qquad {}_{\alpha}q\,(t_1, t_2) = 1 - e^{-\int_{t_1}^{t_2} {}_{\alpha}\mu\, dt}$$

risulta:

$$-\,{}_{\alpha}\mu = \frac{log.\ {}_{\alpha}p\,(t_1, t_2)}{(t_2 - t_1)}\,, \qquad /36/$$

si ha:

$${}_{\alpha}q\,(t, t_2) = 1 - e^{\frac{t_2 - t}{t_2 - t_1} log.\ {}_{\alpha}p(t_1, t_2)}$$

e quindi:

$${}_{\alpha}p\,(t, t_2) = {}_{\alpha}p\,(t_1, t_2)^{\frac{t_2 - t}{t_2 - t_1}}\,.$$

Da questa segue:

$$E\,(t_2) = E\,(t_1)\,p\,(t_1, t_2) + k \int_{t_1}^{t_2} {}_{\alpha}p\,(t_1, t_2)^{\frac{t_2 - t}{t_2 - t_1}}\, dt\,, \qquad /37/$$

ma poichè

$$\frac{d}{dt}\,{}_{\alpha}p\,(t_1, t_2)^{\frac{t_2 - t}{t_2 - t_1}} = -\,\frac{log.\ {}_{\alpha}p\,(t_1, t_2)}{t_2 - t_1}\,{}_{\alpha}p\,(t_1, t_2)^{\frac{t_2 - t}{t_2 - t_1}}\,,$$

si può, nella /37/, sostituire alla quantità sotto il segno d'integrale, la espressione che si ricava per essa dall'ultima uguaglianza. Eseguita tale sostituzione, integrando, si ha subito:

$$E\,(t_2) = E\,(t_1)\,{}_{\alpha}p\,(t_1, t_2) - \frac{{}_{\alpha}q\,(t_1, t_2)\,.\,(t_2 - t_1)\,k}{log.\ {}_{\alpha}p\,(t_1, t_2)}\,. \qquad /38/$$

È questa la formula scritta dal Wittstein per ${}_{\alpha}\mu = cost.$ $I\,(t) = cost.$, $U\,(t) = cost.$, $t_1 = 0$, $t_2 = 1$ [1] ed è formula corretta nella stessa ipotesi che rende corretta la /19/, ma poco utile. Il Wittstein ne dette una soluzione approssimata ch'egli riteneva soddisfacente per la pratica, ma questa opinione non fu poi

---

[1] T. Wittstein, *Die Mortalität in Gesellschaften mit successiv eintretenden und ausschreidenden Mitgliedern*, in: « Grunert's Archiv der Mathematik und Physik », Bd. 39, 1862, pag. 77, formula /25/.

condivisa dallo Zeuner (1), e la ipotesi ${}_{\alpha}\mu = cost.$ passò tra quelle ipotesi che portano ad equazioni risolvibili soltanto per successive approssimazioni (2), e fu lasciata definitivamente in disparte.

Ma se anche il Wittstein, o chi per esso, avesse dedotto la /35/ dalla classica /19/, com'è certo che si possa fare (3), ma non era stato fatto prima del Gini, in verità avrebbe dimostrato qualche cosa meno della /35/, perchè avrebbe vincolato questa formula alle stesse condizioni di validità della /19/, e quindi avrebbe limitato il suo effettivo campo di applicazione. Infatti, la ${}_{\alpha}\mu = \text{cost.}$ si riferisce al complesso degli $E(t)$ nei successivi intervalli di tempo, ma non esclude che i singoli gruppi di questi (sopravviventi dei successivi gruppi di entrati, nuovi entrati e anche usciti) abbiano, separatamente, saggi istantanei diversi, e quindi diverse leggi pure di eliminazione. Per convincersene, ove fosse bisogno, facciamo un esempio semplicissimo.

Da una collettività di 100000 individui sono, in un primo intervallino di tempo, eliminati, per la causa α, 20 individui: il tasso di eliminazione, ragguagliato a unità di tempo, è dunque $20/_{100000\Delta t'}$ dove Δt è l'ampiezza dell'intervallino. Nell'intervallino successivo, dei 100000 esposti iniziali restano in osservazione soltanto 90000 individui, 9980 essendo eliminati per altre cause. Da questi provengono 27 eliminati per α, e quindi il tasso relativo è $27/_{90000\ \Delta t} = 3/2$ del precedente. Se supponiamo che nel 2° in-

---

(1) G. Zeuner, *Abhandlungen aus der mathematischen Statistik*, Leipzig, 1869, pag. 110. Osservo, come curiosità, che la formula riportata dallo Zeuner, differisce, per errore, da quella del Wittstein, perchè si scrive, coi nostri simboli

$$E(t_2) = {}_{\alpha}p\ (t_1\ ,\ t_2) \left\{ E(t_1) - \frac{{}_{\alpha}q\ (t_1\ ,\ t_2)\ k}{\log.\ {}_{\alpha}p\ (t_1\ ,\ t_2)} \right\}$$

e collo stesso errore la /38/ è riportata anche nel *Sunto dei saggi di statistica matematica*, pubblicato dalla « Direzione Generale della Statistica Italiana nel 1883 » (pag. 58, formula /22/).

(2) G. Balducci, *Costruzione e critica delle tavole di mortalità*, in: « Studi di Economia Finanza e Statistica » editi dal Giornale degli Economisti e Rivista di Statistica, Athenaeum, Roma, 1917.

(3) F. P. Cantelli, *Sulla costruzione » delle tavole, di mortalità*, in: « Atti della III riunione della S.I.P.S. », Giugno-Luglio 1941.

tervallino sieno aggiunti ai precedenti altri 110000 individui, i quali forniscono 13 eliminati, si ha un tasso complessivo di $(27+13)/_{(90000+110000)\,\Delta t} = 40/_{200000\,\Delta t}$ cioè propio uguale al precedente valore.

L'esempio si presta a una obbiezione. E questa è che noi non abbiamo, in effetti, calcolato i tassi di eliminazione relativi a due istanti successivi, bensì le probabilità di eliminazione relative a due intervallini successivi, e abbiamo mostrato che la somma di gruppi non omogenei può presentare probabilità di eliminazione costanti nel tempo. Era ovvio. Ma i tassi ? Non mi pare difficile imbastire un esempio nel continuo, ma mi sembra superfluo, perchè la obbiezione nasce da un equivoco, che scambia la funzione del tasso istantaneo di eliminazione ${}_a\mu(t)$, pura finzione matematica, con quella del valore $m\,(t,\, t+\Delta t)\,/\,l\,(t)$, probabilità di eliminazione nell'intervallino Δt, che è la quantità concreta dalla quale quel concetto trae origine e giustificazione. Sempre che il 2° rapporto si possa pensare determinabile in intervalli sempre più piccoli, comunque piccoli, si può pensare alla esistenza di un limite, non del rapporto stesso, perchè il numero $m\,(t,\, t+\Delta t)$, necessariamente finito, deve tendere a zero con Δt, ma del quoziente $\dfrac{m\,(t,t+\Delta t)\,/\,l(t)}{\Delta t}$, che esprime la probabilità di eliminazione nell'intervallo Δt, ragguagliata all'unità di tempo, cioè la frequenza degli eliminati tra t e $t+1$, rispetto agli $l\,(t)$ esposti all'inizio dell'intervallo, nella ipotesi della uniforme distribuzione degli eliminati. Si può pensare che $\mu\,(t)$ esista, e questo simbolo è comodo, perchè ci concede il sussidio di quel potente strumento d'indagine che è l'analisi infinitesimale, ma non cessa per questo di essere un simbolo, il valore teorico approssimato di un valore concreto che è il solo che si possa e che si deve determinare nei casi concreti, il solo che abbia un senso . Perchè è un risultato che sembra bene acquisito alla logica scientifica moderna, che non ha senso il concetto che non può essere oggetto di esperienza.

Dunque la obbiezione non è legittima, perchè quando parliamo di tassi istantanei costanti in realtà noi pensiamo a probabilità costanti, relative a intervallini Δt comunque piccoli, ma di necessità finiti.

La /35/ è pertanto davvero una « novità sostanziale », come la definii alla XXIV riunione della S. I. P. S., [1] e ha una importanza che va molto al di là del fatto che costituisce una rettifica impensata della soluzione di un problema posto, e mal risolto, 73 anni prima, da un altro Autore.

Risulta da quanto abbiamo detto, che per determinare la probabilità indipendente di eliminazione, nel caso più generale di una collettività aperta, la sola formula legittima è la /27/

$$q(t_1, t_2) = 1 - e^{-\int_{t_1}^{t_2} \mu(t)\, dt}$$

Ma questa è formula scomoda, perchè $\mu(t)$ è una funzione incognita, ed è incognita perchè non esiste, perchè in realtà il rapporto $\frac{m(t, t+\Delta t)}{E(t)\Delta t}$ per ogni Δt sufficientemente piccolo ha valore zero se nessun individuo della collettività viene eliminato nell'istante preciso t, oppure m/ε ove m è un numero intero e ε una quantità che tende a zero con Δt, se m individui vengono eliminati nell'istante preciso t. Senza le ipotesi irreali degli individui che vengono eliminati pezzetto per pezzetto, $\mu(t)$ è una funzione nulla in tutti i punti dell'intervallo di definizione, salvo un insieme di punti d'infinito, di dimensione nulla.

Ammessa però la continuità della $\mu(t)$, un valore risulta certo per $\mu(t)$, ed è quello definito dalla /28/ $s = \frac{\int \mu(t) E(t) dt}{\int E(t)\, dt,}$, che esprimendo s come media interna dei $\mu(t)$, assicura anche della esistenza di un punto t_0 nel quale è $\mu(t_0) = s(t_1, t_2)$. Poichè questa relazione è valida in ogni intervallo Δt, comunque piccolo, e poichè in ogni intervallo sufficientemente piccolo si può fissare ad arbitrio la legge di distribuzione degli eliminati, quindi la fun-

---

[1] V. Castellano, *Studi e progressi nel campo della metodologia statistica durante gli anni* 1935/36, in : « Atti della XXIV riunione della S. I. P. S. », vol. 2°, fasc. 1°, marzo 1936. Respingo oggi ancora più decisamente di allora la nota che trovai apposta a detta relazione, che avrebbe dovuto dire, e non risulta bene dal testo, che alla mia pretesa di presentare come novità il risultato del Gini mi fu obbiettato, dal prof. Messina e dalla dott.sa M. Castellani che la /35/ doveva, « se esatta » necessariamente dedursi dalla /19/.

zione $\mu(t)$, nulla vieta di assumere per definizione di $\mu(t)$ proprio la $\mu(t) = s(t)$, in ogni intervallo Δt sufficientemente piccolo, e quindi trasformare la /27/ nella

$$q(t_1, t_2) = 1 - e^{-\Sigma s(\Delta t_i)} \qquad /39/$$

che è formula rigorosa se la massima ampiezza degl'intervalli Δt_i non supera la distanza di tempo tra due eliminazioni successive, e tanto più approssimata quanto meno $\mu(t)$, $E(t)$ variano in ciascuno intervallo (formula /31/), se la condizione precedente non è soddisfatta.

Alla /39/ può essere sostituita con molto vantaggio pratico la /32/

$$q(t_1, t_2) = 1 - e^{-(t_2 - t_1) s_t},$$

se $\mu(t)$ o $E(t)$ possono ritenersi approssimativamente costanti $(\bar{E}/_{\alpha}\bar{E} = 1)$ in tutto l'intervallo t_1, t_2.

§ 5. — A titolo esemplificativo proponiamoci di determinare le probabilità di concepimento, tra le coniugate delle famiglie di cui abbiamo discorso avanti. Il materiale da me elaborato differisce da quello del Gini per la inclusione di un nuovo gruppo di famiglie mediatizzate tedesche, che porta a n. 771 il totale delle famiglie considerate. Ho escluso dal computo una famiglia per la quale o i dati risultavano affetti da errore o il primogenito derivava da concepimento antenuziale.

Per la determinazione della fecondità ho adottato il metodo *c*) del Gini [1], che consiste, come abbiamo già visto (pag. 30), nel tener conto delle durate effettive di esposizione al rischio del concepimento, considerando quindi anche escluse, per un certo periodo, dal rischio del concepimento, le donne che abbiano già concepito [2]. Il Gini assume l'anno come ampiezza di tale pe-

---

[1] C. Gini: *Su la curva della fecondità matrimoniale della donna secondo l'età*, già cit.

[2] Il Livi, (*Trattato di Demografia*, vol. II, *Le leggi naturali della popolazione*, pagg. 80-1, Padova, Cedam, 1940) trova errato tale procedimento perchè «si viene a tener conto due volte, nei singoli rapporti, delle variazioni della fecondità. Se ne tiene conto una volta apponendo al numeratore il numero dei parti o dei concepimenti, un'altra volta se ne

riodo di eliminazione temporanea, salvo ad assumerne uno effettivamente più breve in quei casi eccezionali nei quali un parto è seguito al precedente a distanza minore di un anno. Dato lo

---

tiene conto escludendo dal denominatore il numero delle donne impegnate nella procreazione. I coefficienti di fecondità così calcolati sono quindi dei rapporti tra il numero dei concepimenti e il numero delle donne anno in stato di riposo ». L'affermazione può sembrare curiosa, ma a me pare piuttosto espressione di quel profondo disagio nel quale la estensione fatta alla leggera (a chi risale la prima colpa ?) del concetto di probabilità come rapporto di composizione al concetto di probabilità come rapporto di derivazione, ha messo quanti – e si tratta spesso di valorosi cultori – cercano di vedere più a fondo nelle vecchie questioni. Nella nostra particolare questione della probabilità dei concepimenti secondo l'età mi pare risolutivo l'artificio di assumere una unità di tempo più breve del periodo λ (suppongo per semplicità λ uguale a un numero intero di mesi $\geqslant 9$) d'impossibilità susseguente a ogni concepimento seguito da parto, ad esempio il mese. Immagino, ad es., un gruppo di 1000 donne sposate tutte al compimento del 20° anno di età, e tutte di fecondità massima, talché si abbiano 1000 concepimenti all'inizio del 1° mese di matrimonio, 1000 all'inizio del $\lambda + 1^{c}$ mese, 1000 all'inizio del $2\lambda + 1^{o}$, e così via. Non eliminando le spose alla fine di ciascun mese nel quale avvengono i concepimenti, per i restanti $\lambda - 1$ mesi nei quali un concepimento è impossibile, si ottengono i quozienti di fecondità segnati nella col. 2, mentre eliminandole si ottengono le cifre segnate nella col. 4, nella quali i punti interrogativi significano indeterminazione, dato che i rapporti corrispondenti hanno la forma 0/0. Una legittimissima interpolazione conduce dai valori della (4) a quelli della (5), che sono in perfetto accordo con quelli che si otterrebbero interpolando i valori della (6), relativi allo stesso gruppo di donne, nella ipotesi che l'età al matrimonio sia di 20 anni e 1 mese anzichè di 20 anni. Le cifre della col. (3), che si ottengono per la età al matrimonio di 20 anni e 1 mese, col procedimento della non eliminazione temporanea delle donne che hanno concepito, contraddicono invece a quelle della col. 2.

età		quozienti di fecondità				
anni	mesi	(2)	(3)	(4)	(5)	(6)
20	0 / 1	1	—	1	1	?
»	1 / 2	0	1	?	1	1
..	.....	...	...	...	...	...
»	$\lambda - 1 / \lambda$	0	0	?	1	?
»	$\lambda / \lambda + 1$	1	0	1	1	?
»	$\lambda + 1 / \lambda + 2$	0	1	?	1	1
..	.....	...	...	...	...	...
»	$2\lambda - 1 / 2\lambda$	0	0	?	1	?
»	$2\lambda / 2\lambda + 1$	1	0	1	1	?
»	$2\lambda + 1 / 2\lambda + 2$	0	1	?	1	1
..	.....	...	...	...	...	...

scopo esclusivamente metodologico di questa esemplificazione, abbiamo creduto di dover togliere ogni carattere di eccezionalità a simili casi, considerando impossibile un concepimento solo per 11 periodi mestruali dopo il concepimento seguito da parto, e abbiamo eliminato le madri quindi per soli 848‰ di anno dopo ciascun concepimento. Inoltre, avendo constatato che in alcuni casi la morte della donna, essendo intervenuta alla stessa data del parto, o pochi giorni dopo, doveva attribuirsi al parto stesso, e non ad altra causa, abbiamo considerato, quando ciò aveva importanza, la eliminazione per morte come avvenuta alla data del reingresso dopo la eliminazione temporanea dovuta al parto.

Diamo anzitutto uno sguardo al nostro materiale. In ogni anno di età sono esposte al rischio del concepimento due masse ben distinte di donne : le nuove spose, che entrano in osservazione nel corso dell'anno, e le donne sposate negli anni di età precedenti e non eliminate per altre cause, che sono in osservazione all'inizio dell'anno di età, e per tutta la sua durata, a meno che non intervengano cause di eliminazione (scioglimento del matrimonio, fine del periodo statistico di osservazione). L'intervento di tali cause di eliminazione è però poco frequente per le età giovanili (20-30 anni) durante le quali avviene il maggior numero d'ingressi per matrimonio, mentre aumenta nelle età più avanzate, nelle quali diventa minimo il numero dei nuovi ingressi per matrimonio. La probabilità di concepire è molto diversa per le donne di questi due gruppi, in relazione alla diversa durata media del matrimonio.Una idea della differenza può essere fornita dal confronto tra la probabilità di concepire nello stesso anno di età nel quale è avvenuto il matrimonio, e la probabilità di concepire in ciascun successivo anno di età. La prima, ${}_{i}q_{t}$. è espressa dal rapporto tra il numero ${}_{i}n_{t}$ dei concepiti nell'anno dalle i_{t} donne sposate nello stesso anno di età $t, t+1$, ${}_{i}q_{t} = {}_{i}n_{t}/i_{t}$; mentre della seconda possiamo calcolare un valore approssimato per eccesso eliminando al principio dell'anno le donne che risultano eliminate per altre cause nel corso di ciascun anno di età. Indicato con n_{t} il numero di tutti i concepimenti dell'anno di età $t, t+1$, e con u_{t} il numero degli egressi, tale valore ${}_{i}q_{t}$ si ootterrà ragguagliando il numero $n_{.} - {}_{i}n_{t}$ dei concepimenti di donne sposate in età minore di t, al numero $\sum_{r}^{t-1} i_{r} - \sum_{r}^{t} u_{r}$, che è certamente minore del numero effettivo delle esposte al rischio.

Il calcolo, eseguito distintamente per il gruppo delle famiglie sovrane e mediatizzate tedesche, che comprendono complessivamente 391 famiglie con 1212 figli nati vivi (n^o medio di figli $=3{,}10$) e per altre famiglie principesche, ammontanti in complesso a n. 380, con 970 figli nati vivi (n.° medio di figli $= 2{,}55$) dà i risultati molto chiari, che riporto nella tav. II.

Poichè però si potrebbe obbiettare che mentre tutte le nuove spose rappresentano casi possibili per il concepimento, in quanto nel nostro materiale non esistono casi di concepimenti antenuziali, le altre donne non costituiscono tutte casi possibili, in quanto parte di esse, per aver concepito l'anno precedente, non possono più concepire, almeno per un certo periodo di tempo, abbiamo calcolato anche i saggi di eliminazione, tenendo conto, per le nuove spose, delle durate effettive anno di osservazione, e per le altre, di una valutazione largamente errata per difetto di tali durate, ottenuta considerando casi impossibili per tutto l'anno di età $t+1$, $t+2$ le donne che abbiano concepito durante l'anno t, $t+1$.

I valori:

$$ {}_{i}s_t = \frac{{}_{i}n_t}{{}_{i}i_t - {}_{i}e_t} \quad , \quad {}_{k}s_t = \frac{n_t - {}_{i}n_t}{\sum_{r=1}^{t-1} i_r - \sum_{r}^{t} n_r - n_{t-1} - (n_t - {}_{i}n_t)/2} \quad , $$

ove ${}_{i}e_t$ rappresenta la somma delle frazioni di anno trascorse dal compimento del t. esimo anno di età alla data del matrimonio, sono esposti nella tav. III, e mostrano una divergenza in complesso anche maggiore che non quelli della tav. II. La inversione del senso della disuaguaglianza, che spesso si constata per le età fino a 21 anni, stante i risultati della tav. II, è da ascriversi alla larga approssimazione adottata, e alle irregolarità dipendenti dal ristretto numero delle nuove spose (specie per le età 16-18).

Per le età nelle quali si ha un numero non troppo piccolo di nuove spose, i valori di ${}_{i}s_t$ sono press'a poco doppi dei valori di ${}_{i}q_t$ in quanto i matrimoni si distribuiscono nel corso dell'anno in maniera quasi uniforme: ${}_{i}s_t$ ha quindi il significato di probabilità di un concepimento entro un periodo di 1/2 anno di durata del matrimonio, ragguagliata a anno, cioè moltiplicata per 2. Non rappresenta pertanto nè una illogicità, nè un inconveniente il fatto che assuma a volte valori > 1, e che per l'età 16-17 rag-

Tav. II.

Anni di età	Famiglie nobili italiane, francesi, inglesi, austriache, russe, ungheresi, e polacche						Case sovrane di Europa e Case mediatizzate tedesche						Complesso	
	i_t	u_t	$_in_t$	$_iq_t$	$n_t - _in_t$	hq_t	i_t	u_t	n_t	$_iq_t$	$n_t - _in_t$	hq_t	$_iq_t$	hq_t
16–17	3	—	2	0,667	—	—							0,667	
17–18	6	—	2	0,333	1	0,333	7	—	4	0,571	—		0,461	0,333
18–19	22	—	9	0,409	3	0,333	22	—	10	0,454	2	0,286	0,432	0,312
19–20	42	1	25	0,595	9	0,300	29	1	13	0,448	13	0,464	0,535	0,379
20–21	47	4	18	0,383	24	0,353	43	—	22	0,512	23	0,404	0,444	0,376
21–22	55	1	34	0,618	41	0,360	38	1	24	0,632	32	0,323	0,624	0,343
22–23	33	1	14	0,424	48	0,286	28	1	13	0,464	49	0,360	0,443	0,319
23–24	22	5	10	0,454	61	0,311	27	1	13	0,481	54	0,331	0,469	0,320
24–25	28	1	16	0,571	65	0,300	33	4	19	0,576	60	0,323	0,574	0,310
25–26	25	4	9	0,360	54	0,224	27	2	20	0,741	76	0,350	0,558	0,284
26–27	14	5	4	0,286	62	0,237	25	3	10	0,400	70	0,290	0,359	0,236
27–28	12	5	4	0,333	59	0,218	21	5	9	0,429	71	0,272	0,394	0,245
28–29	8	2	4	0,500	46	0,164	16	5	4	0,250	80	0,289	0,333	0,226
29–30	9	4	3	0,333	42	0,148	44	2	7	0,500	71	0,244	0,435	0,197
30–31	12	2	1	0,083	49	0,168	10	5	3	0,300	69	0,230	0,182	0,200
31–32	10	7	6	0,600	43	0,145	7	8	1	0,143	50	0,165	0,412	0,156
32–33	7	7	2	0,286	29	0,097	8	4	2	0,250	57	0,187	0,267	0,142
33–34	4	7	1	0,250	24	0,080	12	4	5	0,417	43	0,141	0,375	0,111
34–35	3	2	—	0 —	34	0,113	5	7	—	0 —	43	0,138	—	0,126

giunga il valore di 5,291 : in effetti, i tre matrimoni celebrati in questa età avevano, alla fine dell'anno, una durata media di 1 mese e mezzo, quindi ${}_{i}s_{16}$ rappresenta la probabilità di concepire in un tempo medio di 1,5 mesi dopo il matrimonio, ragguagliata ad anno, cioè moltiplicata per circa 8.

TAV. III.

Anni di età	Famiglie nobili italiane francesi, inglesi, austriache, russe, ungheresi e polacche		Case sovrane di Europa e Case mediatizzate tedesche		Complesso	
t	${}^{i}s_{t}$	${}^{h}s_{t}$	${}^{i}s_{t}$	${}^{h}s_{t}$	${}^{i}s_{t}$	${}^{h}s_{t}$
16–17	5,291	—	—	—	5,291	—
17–18	0,590	2,000	1,177	—	0,884	2,000
18–19	0,761	0,667	0,986	1,000	0,865	0,769
19–20	1,138	0,667	0,865	1,368	1,027	0,957
20–21	0,809	1,091	1,097	1,179	0,969	1,133
21–22	1,184	0,796	1,286	0,842	1,224	0,816
22–23	0,882	0,696	0,974	0,883	0,924	0,779
23–24	0,864	0,589	1,065	0,730	0,967	0,648
24–25	0,974	0,573	1,056	0,674	1,017	0,617
25–26	0,680	0,406	1,170	0,760	0,956	0,558
26–27	0,624	0,371	0,755	0,636	0,712	0,477
27–28	0,601	0,338	0,846	0,488	0,752	0,406
28–29	0,949	0,237	0,613	0,510	0,745	0,359
29–30	0,539	0,197	0,948	0,414	0,773	0,294
30–31	0,318	0,221	0,634	0,368	0,508	0,289
31–32	0,163	0,192	0,268	0,244	0,788	0,217
32–33	0,611	0,123	0,606	0,253	0,608	0,187
33–34	0,486	0,094	0,832	0,191	0,744	0,139
34–35	—	0,131	—	0,178	0, —	0,154

Per il solo fatto della maggiore fecondità delle nuove spose il metodo delle durate esatte risulta illegittimo ai fini della determinazione della probabilità di un concepimento. Con ciò non si afferma che questo metodo non possa fornire, in alcuni anni di età, dei risultati anche buoni, il che potrebbe verificarsi, specie per le età più elevate, oltre i 30 anni, in quanto in queste si attenua l'importanza dei nuovi matrimoni, ed è ipotesi plausibile, ad ogni modo non sconfessabile, che le spose eliminate per cause diverse dal concepimento non costituiscano un gruppo scelto rispetto alla massa delle coetanee ; nè si nega che possa fornire qualche valore esatto, o affetto da piccolo errore, anche per età minore dei 30 anni, dato che l'eventuale errore che si commette colla ipotesi $q_t = (t + 1 - u) \cdot q\,(u, t + 1)$ potrebbe anche compensare in

parte l'errore dovuto alla illegittimità del metodo. Ciò è altra cosa: si afferma solo che la rispondenza o meno, in pratica, dei risultati, va giustificata con la concordanza coi risultati di altri metodi più legittimi.

Se chiudiamo il gruppo delle esposte in ciascun anno di età, prendendo in osservazione le donne sposate nell'anno di età t, $t + 1$ all'inizio del successivo anno $t + 1, t + 2$; riprendendole in osservazione, dopo un concepimento, all'inizio dell'anno di età successivo a quello nel quale ha termine il periodo ($\lambda = 0{,}848$) di eliminazione temporanea per impossibilità di concepire, dopo un precedente concepimento; ed eliminandole definitivamente all'inizio dell'anno nel quale ha termine il periodo di osservazione; – e detraiamo, s'intende, dal numero complessivo dei concepimenti quelli avvenuti durante i periodi, così ampliati, di esclusione dall'osservazione – i gruppi chiusi così ottenuti presenteranno una fecondità inferiore a quella dei gruppi aperti reali, oltre che per la esclusione delle nuove spose, anche per la più prolungata esclusione delle spose più feconde. Infatti, le spose sterili, non avendo concepito restano escluse dalla osservazione solo per il periodo che intercorre tra il matrimonio e il compimento dell'anno di età nel quale sono sposate, in ogni caso, per meno di un anno; le spose che hanno concepito una prima volta sono eliminate per tutto l'anno nel quale è avvenuto il concepimento, se questo è avvenuto in età tra $t, t + 1 - \lambda$, e anche per tutto l'anno seguente, se è avvenuto nella successiva frazione $t + 1 - \lambda, t + 1$ dell'anno di età $t, t + 1$. Se nell'intervallo interviene un altro concepimento le donne sono riammesse in osservazione allo scadere dell'anno di età nel quale si compie il periodo λ di eliminazione temporanea dopo il 2° concepimento, e così via. Si comprende come le donne più feconde, per le quali i parti sono più ravvicinati, più frequentemente sono escluse dalla osservazione, col risultato di abbassare i quozienti di fecondità. Il gruppo chiuso ha però il grande vantaggio di permettere, attraverso il calcolo di ${}_{e}q_{t}$, la determinazione precisa di $-\int \mu(t)\, dt = log.\,(1 - {}_{e}q_{t})$. Il confronto di $\int \mu\,(t)\, dt$ col valore corrispondente di ${}_{e}s_{t}$ permette di giudicare sulla accettabilità o meno della ipotesi $\mu\,(t) = cost.$, che consente la sostituzione di quest'ultimo valore – sempre esattamente calcolabile, perchè $= {}_{e}s_{t}$ – al precedente, nel gruppo chiuso. Ma poichè il gruppo aperto reale si ottiene aggiungendo il gruppo delle

nuove spose – le cui età d'ingresso sono abbastanza uniformemente distribuite entro ciascun anno di età, come dimostra il risultato precedente ${}_{i}s_{t} \simeq 2\,{}_{i}q_{t}$, e la cui fecondità non varia decisamente coll'età –; le spose eliminate oltre il periodo λ dopo ciascun concepimento – che non presentano una particolare distribuzione nel tempo –; e le spose eliminate prima dell'effettivo termine definitivo della loro osservazione – che non contribuiscono quasi nulla ai concepimenti – è ben fondata la presunzione che, se accordo tra $\int \mu\,(t)\,dt$ e s_t esiste nel gruppo chiuso, anche migliore concordanza deve esistere tra gli stessi valori nel gruppo aperto reale. Tale presunzione è, se non necessaria, certamente utile, in conseguenza del fatto che $\int \mu(t)\,dt$ non è determinabile nel gruppo aperto reale, se non approssimativamente.

In tale gruppo chiuso risultano esposti al rischio del concepimento nell'anno di età $t, t+1$ tutte le donne sposate negli anni di età precedenti, meno quelle eliminate a tutto l'anno di età $t, t+1$ (sempreché non sieno sposate nello stesso anno di età $t, t+1$), meno quelle che hanno concepito in età $t-\lambda, t$, purchè non sieno state eliminate definitivamente tra le età $t, t+1$.

Distinguiamo con uno o con due apici in alto a destra rispettivamente i concepimenti avvenuti nella frazione $t, t+1-\lambda$, e $t+1-\lambda,\ t+1$, dell'anno di età $t, t+1$; indichiamo in basso a destra, con l'età iniziale, l'anno di età cui si riferiscono i simboli; in basso a sinistra il gruppo dei matrimoni o degli usciti, dai quali i concepimenti provengono (talchè il simbolo ${}_{u}n'_{t}$ significa: numero dei concepimenti avvenuti nell'intervallo di età $t, t+1-\lambda$ da donne sposate e eliminate definitivamente durante l'anno di età $t, t+1$; il simbolo ${}_{u_{t-1,t}}n''_{t-1}$ significa: numero dei concepimenti avvenuti durante l'intervallo di età $t-\lambda, t$ da donne eliminate definitivamente entro lo stesso anno di età o nel successivo, e similmente).

Il numero $\bar{\bar{n}}_t$ dei concepimenti delle donne del gruppo chiuso, in età $t, t+1$, risulta espresso dalla relazione:

$$\bar{\bar{n}}_t = \bar{n}'_t + \bar{n}''_t = n_t - ({}_{i}n_t + {}_{u}n_t - {}_{iu}n_t + {}_{n''}n_t - {}_{un''}n_t) \qquad /40/$$

nella quale ${}_{n''}n_t$ rappresenta il numero dei concepimenti provenienti da donne che abbiano già concepito nell'intervallo di età

$t - \lambda, t$; e $_{w}n''_{t}$ indica il numero dei concepimenti, fra i precedenti, provenienti da donne eliminate definitivamente entro l'anno di età t, $t + 1$.

Il numero $_{c}E_{t}$ delle donne esposte al rischio risulta:

$$_{c}E_{t} = \sum_{r}^{t-1} i_{r} - (\sum_{r}^{t} u_{r} - {}_{i}u_{t}) - (n''_{t-1} - {}_{u}n''_{t-1}) \qquad /41/$$

Il numero degli anni di durata di esposizione di queste donne si otterrà togliendo agli anni di durata $_{c}E_{t}$ una frazione λ di anno per ciascuna delle $\bar{n}'_{t}$ donne che hanno concepito in età $t, t + 1 - \lambda$, e le frazioni d'anno successive al concepimento per ciascuna delle $\bar{n}''_{t}$ donne che concepiscono tra le età $t + 1 - \lambda$ e $t + 1$. Se $_{\bar{n}''}e_{t}$ indica la somma delle frazioni d'anno intertercorrenti, per ciascuna di queste donne, tra l'inizio dell'anno di età e il concepimento, la somma degl'intervalli di non esposizione al rischio sarà espressa dalla quantità:

$$_{c}S_{t}(\lambda) = \lambda \cdot \bar{n}'_{t} + (\bar{n}''_{t} - {}_{\bar{n}''}e_{t}), \qquad /42/$$

e quindi, in definitiva, i valori di $_{c}q_{t}$ e $_{c}s_{t}$ saranno espressi dalle formule:

$$_{c}q_{t} = \frac{\bar{n}_{t}}{{}_{c}E_{t}} \qquad /43/ \qquad {}_{c}s_{t} = \frac{\bar{n}_{t}}{{}_{c}E_{t} - {}_{c}S_{t}(\lambda)} \qquad /44/$$

Questi valori sono riportati, per entrambi i gruppi considerati, e per il gruppo totale, nella tabella IV, dalla quale risultano anche le approssimazioni della uguaglianza $q_{t} \cong 1 - e^{-s_{t}}$. Dalla tabella risulta che la sostituzione del 1° membro di questa ultima uguaglianza approssimata al 2° membro si può ritenere corretta per le età dai 30 anni in poi, soddisfacente dai 20 ai 30, insoddisfacente per le età minori di 20 anni.

Per il calcolo delle durate di esposizione al rischio del gruppo aperto reale, occorre tener presente che ove la eliminazione definitiva segua a distanza minore di λ da un concepimento, la eliminazione per il concepimento da temporanea diventa definitiva. Ciò si otterrà escludendo dalle eliminazioni le $_{n\lambda}u_{t}$ donne eliminate defini-

tivamente a distanza $<\lambda$ da un concepimento, e aggiungendo le ${}_{u}{}_{\lambda}n_t$ donne che hanno concepito a distanza $<\lambda$ dalla eliminazione definitiva. Questi numeri possono eventualmente coincidere, ma in generale non coincideranno, le ${}_{n}{}_{\lambda}u_t$ eliminazioni dell'anno di età $t, t+1$ potendo seguire a concepimenti dello stesso anno o dell'anno precedente, mentre gli ${}_{u}{}_{\lambda}n_t$ possono essere seguiti da eliminazione nello stesso anno di età e nell'anno successivo.

Posto

$$\overset{t}{\Sigma} u'_r = \overset{t}{\Sigma} u_r - {}_{n}{}_{\lambda}u_t + {}_{u}{}_{\lambda}n_t, \qquad /45/$$

e corrispondentemente

$${}_{u}e_t' = {}_{u}e_t - {}_{un\lambda}e_t + {}_{nu\lambda}e_t \qquad /46/,$$

la somma delle durate di esposizione nell'anno di età t, $t+1$, non tenendo conto della durata di eliminazione temporanea, risulta:

$${}_{s}D_t = (\overset{t}{\Sigma} i_r - {}_{s}e_t) - (\overset{t}{\Sigma} u'_r - {}_{u}e'_t) \qquad /47/$$

La somma ${}_{s}S_t(\lambda)$ delle durate di eliminazione temporanea che cadono nell'anno di età $t, t+1$, è composta di due parti: $S_{t-1}(\lambda)$, relativa agli n''_{t-1} concepimenti dell'anno precedente, per i quali la fine del periodo di eliminazione temporanea cade nell'anno di età $t, t+1$, e $S_t(\lambda)$ relativa ai concepimenti dell'anno in corso. Osserviamo che se la durata della eliminazione temporanea fosse di 1 anno, ciascuna madre rientrerebbe in osservazione esattamente alla stessa età, cioè alla stessa età precisa del concepimento, più 1 anno, e quindi tutte dovrebbero essere eliminate, complessivamente, nell'anno di età seguente, per una durata ${}_{n}e_{t-1}$. Il periodo di eliminazione essendo di $1-\lambda$ più breve dell'anno, non danno luogo a eliminazione nell'anno successivo le n'_{t-1} donne che concepiscono tra le età $t-1, t-\lambda$, mentre le n''_{t-1} che concepiscono tra le età $t-\lambda, t$, dovranno essere eliminate per una durata $(1-\lambda)\, n''_{t-1}$ minore della ${}_{n''}e_{t-1}$ che loro competerebbe se la durata di eliminazione fosse di un intero anno. Risulta pertanto :

$$S_{t-1}(\lambda) = {}_{n''}e_{t-1} - (1-\lambda)\, n''_{t-1} \qquad /48/$$

TAV. IV.

Anni di età	Famiglie nobili italiane, francesi, inglesi, austriache, russe, ungheresi e polacche			Case sovrane di Europa e case mediatizzate tedesche			Complesso		
	$c^{s}t$	$c^{q}t$	$1-e^{-s}t$	$c^{s}t$	$c^{q}t$	$1-e^{-s}t$	$c^{s}t$	$c^{q}t$	$1-e^{-s}t$
17–18	4,9261	1,0000	0,993				4,9261	1,0000	0,993
18–19	0,4924	0,3750	0,389	0,4647	0,3333	0,372	0,4851	0,3636	0,384
19–20	0,4813	0,3158	0,382	0,8067	0,5294	0,554	0,6350	0,4167	0,470
17–20	0,5060	0,3571	0,397	0,7514	0,5000	0,528	0,6048	0,4167	0,454
20–21	0,5692	0,4390	0,434	0,5022	0,3939	0,395	0,5390	0,4189	0,417
21–22	0,5201	0,4079	0,406	0,4049	0,3226	0,333	0,4679	0,3696	0,374
22–23	0,4396	0,3271	0,356	0,4497	0,3596	60362	0,4443	0,3418	0,359
23–24	0,4409	0,3542	0,357	0,4295	0,3364	0,349	0,4360	0,3465	0,353
24–25	0,4168	0,3459	0,341	0,5146	0,3923	0,402	0,4588	0,3668	0,368
20–25	0,4540	0,3604	0,365	0,4613	0,3608	0,370	0,4572	0,3596	0,367
25–26	0,2814	0,2471	0,245	0,4913	0,3935	0,388	0,3767	0,3169	0,314
26–27	0,2918	0,2547	0,253	0,3520	0,2975	0,297	0,3170	0,2730	0,272
27–28	0,2731	0,2381	0,239	0,3749	0,3109	0,313	0,3206	0,2730	0,274
28–29	0,1883	0,1703	0,172	0,3649	0,3052	0,306	0,2700	0,2353	0,237
29–30	0,1746	0,1667	0,160	0,3116	0,2615	0,268	0,2363	0,2078	0,210
25–30	0,2363	0,2103	0,210	0,3724	0,3095	0,311	0,2977	0,2567	0,257
30–31	0,1816	0,1653	0,166	0,2749	0,2397	0,240	0,2261	0,2016	0,202
31–32	0,1734	0,1589	0,159	0,1928	0,1757	0,175	0,1827	0,1670	0,167
32–33	0,1138	0,1089	0,108	0,2162	0,1916	0,194	0,1634	0,1506	0,151
33–34	0,0799	0,0769	0,077	0,1714	0,1583	0,157	0,1235	0,1166	0,116
34–35	0,1170	0,1095	0,110	0,1617	0,1487	0,149	0,1386	0,1286	0,129
30–35	0,1313	0,1230	0,123	0,2014	0,1819	0,182	0,1650	0,1518	0,152
35–36	0,0655	0,0637	0,063	0,1282	0,1194	0,120	0,0963	0,0916	0,092
36–37	0,0965	0,0929	0,092	0,1258	0,1181	0,118	0,1110	0,1056	0,105
37–38	0,0797	0,0769	0,077	0,0742	0,0717	0,072	0,0770	0,0744	0,074
38–39	0,0664	0,0637	0,064	0,0961	0,0913	0,092	0,0812	0,0778	0,078
39–40	0,0415	0,0408	0,041	0,0565	0,0547	0,055	0,0491	0,0479	0,048
35–40	0,0703	0,0681	0,068	0,0965	0,0917	0,092	0,0833	0,0800	0,080
40–41	0,0173	0,0171	0,0170	0,0203	0,0201	0,0201	0,0189	0,0186	0,0187
41–42	0,0452	0,0442	0,0433	0,0367	0,0361	0,0360	0,0408	0,0400	0,0400
42–43	0,0046	0,0046	0,0046	0,0340	0,0333	0,0334	0,0198	0,0196	0,0196
43–44	0,0193	0,0191	0,0191	0,0044	0,0044	0,0044	0,0115	0,0115	0,0114
44–45	0, /	0, /	0, /	0,0140	0,0140	0,0139	0,0072	0,0072	0,0072
40–45	0,0176	0,0174	0,0175	0,0223	0,0220	0,0221	0,0200	0,0198	0,0198
45–46	0, /	0, /	0, /	0,0053	0,0053	0,0053	0,0027	0,0027	0,0027
46–47	0, /	0, /	0, /	0,0057	0,0057	0,0057	0,0029	0,0029	0,0029
45–47	0, /	0, /	0, /	0,0055	0,0055	0,0055	0,0028	0,0028	0,0028
17–47	0,1287	0,1203	0,121	0,1649	0,1511	0,152	0,1463	0,1355	0,136

Circa i concepimenti dell'anno basta osservare che tutto il periodo di eliminazione temporanea cade nell'anno stesso per gli n_t' concepimenti che avvengono nella prima frazione $1-\lambda$ dell'anno di età, mentre le n_t'' donne che concepiscono nella residua frazione λ restano eliminate per tutto il resto dell'anno. Non occorre eliminare temporaneamente, s'intende, le donne il cui concepimento è seguito a meno di λ dalla eliminazione definitiva, in quanto esse sono già state eliminate definitivamente al concepimento, per le /45/ e /46/. È quindi:

$$S_t(\lambda) = (n_t' - {}_{u\lambda}n_t')\,\lambda + [(n_t'' - {}_{u\lambda}n_t'') - ({}_{n''}e_t - {}_{n''u\lambda}e_t)] \qquad /49/$$

La somma ${}_aD_t(\lambda)$ delle durate effettive di esposizione, tenuto conto della eliminazione per il tempo λ dopo il concepimento, risulta quindi dalla formula:

$${}_aD_t(\lambda) = {}_aD_t - {}_aS_t(\lambda) \qquad /50/,$$

dove

$${}_aS_t(\lambda) = S_{t-1}(\lambda) + S_t(\lambda) \qquad /51/,$$

e il saggio:

$$s_t = n_t \,/\, {}_aD_t(\lambda) \qquad /52/$$

Per calcolare le durate di esposizione occorrenti per l'applicazione del metodo delle durate esatte, non basta aumentare ${}_aD_t(\lambda)$ di $S_t(\lambda)$, ma occorre anche eliminare alla fine dell'anno di età, anzichè al concepimento, le ${}_{u\lambda}n_t$ donne la cui eliminazione definitiva cade in un intervallo $<\lambda$ dal concepimento, il che si ottiene aggiungendo la quantità ${}_{u\lambda}n_t - {}_{nu\lambda}e_t$, ed eliminando anche alla fine dell'anno, anzichè alla data della eliminazione definitiva le ${}_{n'}u_t - {}_{n'\lambda}u_t$ donne le cui eliminazione definitiva segue ad un intervallo λ il concepimento (avvenuto nello stesso anno), occorre aggiungere ancora il termine $({}_{n'}u_t - {}_{n'\lambda}u_t) - ({}_{un'}e_t - {}_{un'\lambda}e_t)$.

Indicando con $\varepsilon(n, u)$ il complesso di tali due termini correttivi:

$$\varepsilon(n,u) = ({}_{u\lambda}n_t - {}_{nu\lambda}e_t) + ({}_{n'}u_t - {}_{n'\lambda}u_t) - ({}_{un'}e_t - {}_{un'\lambda}e_t) \qquad /53/,$$

si ha, per la quantità ${}_Aq_t$ che esprime la probabilità di un concepimento col metodo delle durate esatte dell'Ackland, la formula:

$${}_Aq_t = \frac{n_t}{{}_aD_t(\lambda) + S_t(\lambda) + \varepsilon(n,u)} \qquad /54/$$

L'applicazione delle formule /52/ e /54/ richiede classificazioni minuziose del materiale, che si possono dedurre tutte dalle fondamentali classificazioni degl'ingressi, degli egressi per la causa in esame e degli egressi per altre cause, per anno di età, come normalmente si eseguono per il metodo delle durate esatte, purchè si usi: l'accorgimento di premettere ai numeri decimali corrispondenti alle relative frazioni degli anni di età, il numero della scheda in esame, nel nostro caso, il numero della madre considerata. Così, per es., se la donna n. 89 del gruppo delle Case sovrane, sposa in età 22,677, concepisce alle età 22,690 ; 23,882 ; 26,047 ; è eliminata definitivamente in età 28,101 ; nella colonna 22-23 degl'ingressi si scriverà 89,677 invece di 0,677 ; nelle colonne 22-23, 23-24, 26-27 dei concepimenti si scriverà rispettivamente 89,690 ; 89,882 ; 89,047 ; invece di 0,690 ; 0,882 ; 0,047 ; e nella colonna 28-29 delle eliminazioni definitive si scriverà 89,101 invece di 0,101. Naturalmente non si terrà affatto conto della parte intera dei numeri scritti, nella esecuzione delle somme ${}_{i}e_{t}$, ${}_{n}e_{t}$, ${}_{u}e_{t}$, intere o parziali che sieno, le parti intere avendo solo il compito di permettere le più minute classificazioni occorrenti (per es. ; ${}_{i}n_{t}$, ${}_{u}n_{t}$, ${}_{iu}n_{t}$, ecc.) e di rendere agevole il controllo e la correzione degli errori, che sono frequenti in elaborazioni così minute.

Come valore di confronto $\int \mu\,(t)\,dt$ abbiamo assunto la somma dei saggi mensili relativi a ciascun anno di età. Per calcolare le esposte al rischio in ciascun mese abbiamo prima fatto una valutazione per difetto di queste mediante la /42/, opportunamente modificata, e poi una valutazione per eccesso prendendo in osservazione ciascuna donna all'inizio del mese di età nel quale cade effettivamente l'ingresso o il reingresso, ed eliminandole definitivamente alla fine del mese nel quale cade la eliminazione definitiva. La media aritmetica delle due determinazioni, dalla quale si sottragga la metà del numero dei concepimenti del mese, equivale al metodo delle durate medie del Gini. I risultati dei calcoli, esposti distintamente per il gruppo delle Case sovrane e mediatizzate tedesche, per il gruppo delle case principesche di Italia, Francia, Inghilterra, Austria, Ungheria, Russia e Polonia, e per il gruppo totale, sono riportati nelle Tavv. V, VI, e VII. I valori $1 - e^{-\Sigma_s t_i}$; e $1 - e^{-s_t}$ che sono posti a raffronto, sono

calcolati mediante la tabella in appendice, coll'approssimazione da questa consentita.

Per i primi due anni di età, 16-17 e 17-18, il numero delle donne esposte al rischio del concepimento è così esiguo (da 0 a 7)

TAV. V.

	EAq_t	s_t	$1-e^{-s_t}$	Σs_{ti}	$1-e^{-\Sigma s_{ti}}$	$\frac{\Sigma s_{ti}\, l_{ti}}{l_{ti}}$	$1-e^{-12\,\frac{\Sigma s_{ti}\, l_{ti}}{\Sigma l_{ti}}}$
16–17	5,291	10,753	099998			2,400	0,909
17–18	0,614	1,253	0,714			1,200	0,700
18–19	0,582	0,817	0,558	0,815	0,557	0,816	0,558
19–20	0,698	1,032	0,644	1,038	0,646	1,015	0,638
20–21	0,535	0,709	0,508	0,675	0,491	0,676	0,492
21–22	0,592	0,807	0,554	0,797	0,550	0,792	0,547
22–23	0,401	0,497	0,392	0,496	0,391	0,492	0,389
23–24	0,380	0,473	0,377	0,461	0,369	0,461	0,369
24–25	0,386	0,475	0,378	0,466	0,372	0,466	0,372
25–26	0,278	0,317	0,272	0,316	0,271	0,317	0,272
26–27	0,270	0,310	0,267	0,305	0,263	0,305	0,263
27–28	0,245	0,283	0,246	0,280	0,244	0,280	0,244
28–29	0,190	0,208	0,188	0,206	0,186	0,206	0,186
29–30	0,166	0,180	0,165	0,180	0,165	0,179	0,164
30–31	0,178	0,197	0,179	0,194	0,176	0,194	1,176
31–32	0,170	0,184	0,168	0,182	0,166	0,182	0,166
32–33	0,108	0,113	0,107	0,111	0,105	0,111	0,105
33–34	0,086	0,089	0,085	0,088	0,084	0,088	0,084
34–35	0,115	0,123	0,116	0,122	0,115	0,122	0,115
35–36	0,065	0,067	0,065	0,067	0,064	0,067	0,064
36–37	0,095	0,099	0,094	0,098	0,094	0,098	0,094
37–38	0,073	0,076	0,073	0,075	0,073	0,075	0,073
38–39	0,060	0,062	0,060	0,061	0,059	0,062	0,059
39–40	0,039	0,039	0,039	0,039	0,039	0,039	0,039
40–41	0,020	0,020	0,020	0,020	0,020	0,020	0,020
41–42	0,042	0,043	0,042	0,043	0,042	0,043	0,042
42–43	0,004	0,004	0,004	0,004	0,004	0,004	0,004
43–44	0,018	0,018	0,018	0,019	0,019	0,019	0,019
44–45	—	—	—	—	—	—	—
45–46	—	—	—	—	—	—	—
46–47	—	—	—	—	—	—	—

che piccole variazioni accidentali nelle durate di esposizione danno luogo a forti variazioni dei quozienti mensili s_{ti}: in un caso il metodo delle durate medie dà un quoziente infinito, ed è evidente la illegittimità del calcolo, che abbiamo dovuto eseguire col metodo

delle durate esatte del Gini. Sono preferibili per tali anni di età, ai valori della col. 7, i corrispondenti della col. 9.

Tutte e due le tavv. V e VI dimostrano all'evidenza che il metodo di Ackland, illegittimo in teoria nel nostro caso, conduce

TAV. VI.

Case sovrane di Europa e Case mediatizzate tedesche

	EA^{q}_{t}	s_t	$1-e^{-s_t}$	Σs_{ti}	$1-e^{-\Sigma s_{ti}}$	$\frac{\Sigma s_{ti} l_{ti}}{\Sigma l_t}$	$1-e^{-\frac{\Sigma s_{ti} l_{ti}}{\Sigma l_t}}$
16–17	—	—	—	—	—	—	—
17–18	1,177	2,626	0,928	—	—	2,461	0,915
18–19	0,768	1,185	0,694	1,211	0,702	1,129	0,677
19–20	0,665	0,977	0,623	0,996	0,631	0,983	0,626
20–21	0,667	0,971	0,621	0,996	0,631	0,964	0,619
21–22	0,550	0,752	0,529	0,759	0,532	0,746	0,526
22–23	0,480	0,606	0,455	0,594	0,448	0,602	0,452
23–24	0,445	0,577	0,439	0,586	0,444	0,578	0,439
24–25	0,429	0,557	0,427	0,557	0,427	0,552	0,424
25–26	0,458	0,591	0,446	0,592	0,447	0,591	0,446
26–27	0,360	0,429	0,439	0,426	0,347	0,429	0,349
27–28	0,332	0,398	0,329	0,398	0,329	0,397	0,328
28–29	0,327	0,391	0,324	0,391	0,324	0,390	0,323
29–30	0,284	0,344	0,291	0,348	0,290	0,341	0,289
30–31	0,254	0,286	0,249	0,234	0,247	0,284	0,247
31–32	0,182	0,198	0,180	0,198	0,180	0,198	0,180
32–33	0,202	0,229	0,205	0,229	0,205	0,229	0,205
33–34	0,161	0,173	0,159	0,173	0,159	0,173	0,159
34–35	0,145	0,156	0,144	0,155	0,144	0,155	0,144
35–36	0,121	0,130	0,122	0,130	0,122	0,130	0,122
36–37	0,127	0,135	0,126	0,135	0,126	0,135	0,126
37–38	0,082	0,086	0,082	0,086	0,082	0,086	0,082
38–39	0,093	0,098	0,094	0,098	0,094	0,098	0,094
39–40	0,069	0,071	0,069	0,072	0,069	0,072	0,069
40–41	0,023	0,023	0,023	0,023	0,023	0,023	0,023
41–42	0,035	0,035	0,034	0,035	0,034	0,035	0,034
42–43	0,032	0,033	0,032	0,033	0,032	0,033	0,032
43–44	0,004	0,004	0,004	0,004	0,004	0,004	0,004
44–45	0,013	0,013	0,013	0,014	0,014	0,014	0,014
45–46	0,001	0,005	0,005	0,005	0,005	0,005	0,005
46–47	0,001	0,006	0,006	0,006	0,006	0,006	0,006

anche a cattive approssimazioni in pratica, salvo che per i saggi non superiori a 0,15 per i quali tutti i metodi sono più o meno equi-

valenti. Inoltre, poichè i valori $1 - e^{-\Sigma s_{ti}}$ risultano sistematicamente minori dei ${}_{EA}q_t$, chi pretendesse calcolare $\int \mu\,(t)\,dt$ come una somma di probabilità relative agl'intervalli parziali Δt – cioè

TAV. VII.

Complesso

	${}_{EA}q_t$	$'_t$	$1-e^{-s_t}$	$\Sigma s_{t.}$	$1-e^{-\Sigma s_{ti}}$	$\frac{\Sigma s_{ti}\, l_{ti}}{l_{t.}}$	$1-e^{-12\,\frac{\Sigma s_{ti}\, '_{ti}}{\Sigma l_{ti}}}$
16–17	5,291	10,753	099998	—	—	2,400	0,910
17–18	0,845	1,787	0,833	—	—	1,697	0,827
18–19	0,662	0,967	0,621	0,970	0,621	0,947	0,612
19–20	0,683	1,007	0,635	1,014	0,637	1,001	0,633
20–21	0,596	0,824	0,561	0,811	0,556	0,802	0,551
21–22	0,574	0,783	0,543	0,779	0,541	0,771	0,538
22–23	0,437	0,546	0,421	0,539	0,417	0,451	0,418
23–23	0,409	0,519	0,405	0,513	0,402	0,511	0,400
24–25	0,406	0,512	0,401	0,506	0,397	0,505	0,397
25–26	0,365	0,440	0,356	0,440	0,356	0,440	0,356
26–27	0,313	0,366	0,307	0,361	0,303	0,363	0,304
27–28	0,287	0,337	0,286	0,336	0,285	0,335	0,285
28–29	0,258	0,294	0,255	0,292	0,254	0,292	0,254
29–30	0,227	0,258	0,227	0,257	0,227	0,256	0,226
30–31	0,216	0,241	0,214	0,238	0,212	0,238	0,212
31–32	0,175	0,191	0,174	0,190	0,173	0,190	0,173
32–33	0,155	0,169	0,156	0,168	0,155	0,168	0,155
33–34	0,124	0,131	0,123	0,130	0,122	0,130	0,122
34–35	0,130	0,139	0,130	0,139	0,130	0,139	0,130
35–36	0,094	0,098	0,093	0,098	0,094	0,098	0,094
36–37	0,111	0,117	0,110	0,117	0,110	0,116	0,110
37–38	0,078	0,081	0,078	0,081	0,078	0,081	0,078
38–39	0,077	0,080	0,077	0,080	0,077	0,080	0,077
39–40	0,054	0,056	0,054	0,054	0,053	0,055	0,054
40–41	0,021	0,022	0,021	0,022	0,021	0,022	0,021
41–42	0,038	0,039	0,038	0,039	0,038	0,039	0,038
42–43	0,019	0,019	0,019	0,019	0,019	0,019	0,019
43–44	0,011	0,011	0,011	0,011	0,011	0,011	0,011
44–45	0,007	0,007	0,007	0,007	0,007	0,007	0,007
45–46	0,002	0,002	0,002	0,003	0,003	0,003	0,003
46–47	0,003	0,003	0,003	0,003	0,003	0,003	0,003

considerando gli eliminati per la causa α eliminati alla fine di ciascun intervallino – eleverebbe alquanto il valore di $1 - e^{-\int \mu(t)dt}$ e quindi farebbe apparire l'approssimazione dei ${}_{AE}q_t$ ancora peggiore

di quella che è in effetti. Invece l'approssimazione delle formule del Gini si rivela abbastanza soddisfacente, specie nella Tav. VII a causa del maggior numero di casi che questa comprende.

Data poi le maggiore semplicità del calcolo della

$$\frac{\Sigma E_{ti}\, s_{ti}}{\Sigma E_{ti}} = \frac{\Sigma m_{ti}}{\Sigma E_{ti}}$$

di fronte alla Σs_{ti}, ho calcolato anche le probabilità annue con la formula

$$q_t = 1 - e^{-12\frac{\Sigma E_{ti}\, s_{ti}}{\Sigma E_{ti}}} \qquad /55/$$

che dà approssimazioni ottime per quasi tutte le età, ed è senza altro da preferire alla più esatta – in generale – $q_t = 1 - e^{-\Sigma s t}$ per le età più giovanili, perchè meno sensibile di questa al piccolo numero delle osservazioni.

APPENDICE: *Tabella della funzione* $s = -\log.(1-q)$ $\quad (q = 1 - e^{-s})$

q	0	1	2	3	4	5	6	7	8	9
0,00	—	0,0010	0,0020	0,0030	0,0040	0,0050	0,0060	0,0070	0,0080	0,0090
0,01	0,0101	0,0111	0,0121	0,0131	0,0141	0,0151	0,0161	0,0171	0,0182	0,0192
0,02	0,0202	0,0212	0,0222	0,0233	0,0243	0,0253	0,0263	0,0273	0,0284	0,0294
0,03	0,0305	0,0315	0,0325	0,0336	0,0346	0,0356	0,0367	0,0377	0,0387	0,0397
0,04	0,0408	0,0419	0,0429	0,0439	0,0450	0,0460	0,0471	0,0481	0,0492	0,0502
0,05	0,0513	0,0523	0,0534	0,0544	0,0555	0,0565	0,0576	0,0587	0,0597	0,0608
0,06	0,0619	0,0629	0,0640	0,0650	0,0661	0,0672	0,0683	0,0693	0,0704	0,0715
0,07	0,0726	0,0736	0,0747	0,0758	0,0769	0,0780	0,0791	0,0801	0,0812	0,0823
0,08	0,0834	0,0845	0,0856	0,0866	0,0877	0,0888	0,0899	0,0910	0,0921	0,0932
0,09	0,0943	0,0954	0,0965	0,0976	0,0987	0,0998	0,1009	0,1020	0,1031	0,1043
0,10	0,1054	0,1065	0,1076	0,1087	0,1098	0,1109	0,1121	0,1132	0,1143	0,1154
0,11	0,1165	0,1176	0,1188	0,1199	0,1210	0,1221	0,1233	0,1244	0,1256	0,1267
0,12	0,1278	0,1289	0,1301	0,1312	0,1324	0,1235	0,1347	0,1358	0,1370	0,1381
0,13	0,1393	0,1404	0,1416	0,1427	0,1439	0,1450	0,1462	0,1473	0,1485	0,1496
0,14	0,1508	0,1520	0,1532	0,1543	0,1555	0,1566	0,1578	0,1590	0,1602	0,1613
0,15	0,1625	0,1637	0,1649	0,1660	0,1672	0,1684	0,1696	0,1708	0,1720	0,1732
0,16	0,1744	0,1755	0,1767	0,1779	0,1791	0,1803	0,1815	0,1827	0,1839	0,1851
0,17	0,1863	0,1875	0,1887	0,1899	0,1912	0,1924	0,1936	0,1948	0,1960	0,1972
0,18	0,1985	0,1997	0,2009	0,2021	0,2033	0,2045	0,2058	0,2070	0,2083	0,2095
0,19	0,2107	0,2119	0,2132	0,2144	0,2157	0,2170	0,2182	0,2194	0,2206	0,2218

APPENDICE: *Tabella della funzione* $s = -log.\ (1-q)$ $\qquad (q = 1 - e^{-s})$

q	0	1	2	3	4	5	6	7	8	9
0,20	0,2231	0,2243	0,2256	0,2269	0,2282	0,2294	0,2307	0,2319	0,2232	0,2344
0,21	0,2357	0,2370	0,2383	0,2395	0,2408	0,2420	0,2433	0,2446	0,2459	0,2472
0,22	0,2485	0,2497	0,2510	0,2523	0,2536	0,2549	0,2562	0,2575	0,2588	0,2601
0,23	0,2614	0,2627	0,2640	0,2653	0,2666	0,2679	0,2692	0,2705	0,2718	0,2731
0,24	0,2744	0,2757	0,2771	0,2784	0,2797	0,2810	0,2824	0,2837	0,2850	0,2863
0,25	0,2877	0,2890	0,2904	0,2917	0,2930	0,2943	0,2957	0,2970	0,2984	0,2997
0,26	0,3011	0,3024	0,3038	0,3051	0,3065	0,3078	0,3092	0,3106	0,3120	0,3133
0,27	0,3147	0,3161	0,3175	0,3188	0,3202	0,3216	0,3230	0,3243	0,3257	0,3271
0,28	0,3285	0,3299	0,3313	0,3327	0,3341	0,3355	0,3369	0,3383	0,3397	0,3411
0,29	0,3425	0,3439	0,3453	0,3467	0,3481	0,3495	0,3510	0,3524	0,3538	0,3552
0,30	0,3567	0,3581	0,3595	0,3610	0,3624	0,3638	0,3653	0,3667	0,3682	0,3696
0,31	0,3711	0,3725	0,3740	0,3754	0,3769	0,3783	0,3798	0,3813	0,3827	0,3842
0,32	0,3857	0,3871	0,3886	0,3901	0,3916	0,3930	0,3945	0,3960	0,3975	0,3990
0,33	0,4005	0,4020	0,4035	0,4050	0,4065	0,4080	0,4095	0,4110	0,4125	0,4140
0,34	0,4155	0,4170	0,4186	0,4201	0,4216	0,4231	0,4246	0,4262	0,4277	0,4292
0,35	0,4307	0,4323	0,4338	0,4354	0,4369	0,4385	0,4400	0,4416	0,4431	0,4447
0,36	0,4463	0,4479	0,4494	0,4510	0,4525	0,4541	0,4557	9,4573	0,4589	0,4005
0,37	0,4620	0,4636	0,4652	0,4668	0,4684	0,4700	0,4716	0,4732	0,4748	0,4764
0,38	0,4780	0,4796	0,4812	0,4829	0,4845	0,4861	0,4877	0,4894	0,4910	0,4927
0,39	0,4943	0,4959	0,4975	0,4992	0,5008	0,5025	0,5041	0,5058	0,5075	0,5092

APPENDICE: *Tabella della funzione* $s = -\log.(1-q)$ $(q = 1 - e^{-s})$

q	0	1	2	3	4	5	6	7	8	9
0,40	0,5108	0,5125	0,5141	0,5158	0,5175	0,5192	0,5208	0,5225	0,5242	0,5259
0,41	0,5276	0,5293	0,5310	0,5327	0,5344	0,5361	0,5378	0,5396	0,5413	0,5430
0,42	0,5447	0,5465	0,5482	0,5499	0,5516	0,5534	0,5551	0,5569	0,5586	0,5604
0,43	0,5621	0,5639	0,5656	0,5674	0,5691	0,5709	0,5727	0,5745	0,5762	0,5780
0,44	0,5798	0,5816	0,5834	0,5852	0,5870	0,5888	0,5906	0,5924	0,5942	0,5960
0,45	0,5978	0,5997	0,6015	0,6033	0,6051	0,6070	0,6088	0,6106	0,6124	0,6143
0,46	0,6161	0,6180	0,6199	0,6218	0,6236	0,6255	0,6273	0,6292	0,6311	0,6330
0,47	0,6349	0,6368	0,6387	0,6406	0,6425	0,6444	0,6463	0,6482	0,6501	0,6520
0,48	0,6539	0,6559	0,6578	0,6597	0,6616	0,6636	0,6655	0,6675	0,6695	0,6714
0,49	0,6733	0,6753	0,6772	0,6792	0,6812	0,6832	0,6852	0,6872	0,6891	0,6911
0,50	0,6931	0,6951	0,6972	0,6992	0,7012	0,7032	0,7052	0,7072	0,7093	0,7113
0,51	0,7134	0,7154	0,7174	0,7194	0,7215	0,7236	0,7257	0,7277	0,7298	0,7319
0,52	0,7340	0,7360	0,7381	0,7402	0,7423	0,7444	0,7465	0,7486	0,7508	0,7529
0,53	0,7550	0,7571	0,7593	0,7614	0,7636	0,7657	0,7679	0,7700	0,7722	0,7743
0,54	0,7765	0,7787	0,7809	0,7831	0,7853	0,7875	0,7897	0,7919	0,7941	0,7963
0,55	0,7985	0,8007	0,8030	0,8052	0,8074	0,8096	0,8119	0,8141	0,8164	0,8187
0,56	0,8210	0,8232	0,8255	0,8278	0,8301	0,8324	0,8347	0,8370	0,8393	0,8416
0,57	0,8440	0,8463	0,8486	0,8509	0,8533	0,8556	0,8580	0,8604	0,8628	0,8651
0,58	0,8675	0,8699	0,8723	0,8747	0,8771	0,8795	0,8819	0,8843	0,8867	0,8891
0,59	0,8916	0,8940	0,8965	0,8989	0,9014	0,9038	0,9063	0,9088	0,9113	0,9138

APPENDICE: *Tabella della funzione* $s = -\log.(1-q)$ $\qquad (q = 1 - e^{-s})$

q	0	1	2	3	4	5	6	7	8	9
0,60	0,9163	0,9188	0,9213	0,9238	0,9263	0,9289	0,9314	0,9339	0,9365	0,9390
0,61	0,9416	0,9441	0,9467	0,9493	0,9519	0,9545	0,9571	0,9597	0,9623	0,9649
0,62	0,9676	0,9702	0,9729	0,9755	0,9782	0,9808	0,9835	0,9862	0.9889	0,9916
0,63	0,9943	0,9970	0,9997	1,0024	1,0051	1,0078	1,0106	1,0133	1,0161	1,0189
0,64	1,0217	1,0244	1,0272	1,0300	1,0328	1,0356	1,0385	1,0413	1,0441	1,0469
0,65	1,0498	1,0527	1,0556	1,0584	1,0613	1,0642	1,0671	1,0700	1,0729	1,0758
0,66	1,0788	1,0817	1,0847	1,0876	1,0906	1,0936	1,0966	1,0996	1,1026	1,1056
0,67	1,1087	1,1117	1,1147	1,1178	1,1209	1,1239	1,1270	1,1301	1,1332	1,1363
0,68	1,1394	1,1425	1,1457	1,1488	1,1520	1,1552	1,1584	1,1616	1,1648	1,1680
0,69	1,1712	1,1744	1,1777	1,1809	1,1842	1,1874	1,1907	1,1940	1,1973	1,2006
0,70	1,2040	1,2073	1,2107	1,2140	1,2174	1,2208	1,2242	1,2276	1,2310	1,2344
0,71	1,2379	1,2413	1,2448	1,2483	1,2518	1,2553	1,2588	1,2623	1,2658	1,2694
0,72	1,2730	1,2765	1,2801	1,2837	1,2874	1,2910	1,2946	1,2983	1,3020	1,3056
0,73	1,3093	1,3130	1,3168	1,3205	1,3243	1,3280	1,3318	1,3356	1,3394	1,3432
0,71	1,3471	1,3509	1,3548	1,3587	1,3626	1,3665	1,3704	1,3743	1,3783	1,3823
0,75	1,3863	1,3903	1,3943	1,3983	1,4024	1,4065	1,4106	1,4147	1,4188	1,4229
0,76	1,4271	1,4313	1,4355	1,4397	1,4439	1,4481	1,4524	1,4567	1,4610	1,4653
0,77	1,4697	1,4740	1,4784	1,4828	1,4872	1,4916	1,4961	1,5006	1,5051	1,5096
0,78	1,5141	1,5187	1,5233	1,5279	1,5325	1,5371	1,5418	1,5465	1,5512	1,5559
0,79	1,5606	1,5654	1,5702	1,5750	1,5799	1,5847	1,5896	1,5945	1,5995	1,6044

q	0	1	2	3	4	5	6	7	8	9
0,80	1,6094	1,6145	1,6195	1,6246	1,6296	1,6347	1,6399	1,6451	1,6503	1,6555
0,81	1,6607	1,6660	1,6713	1,6766	1,6820	1,6874	1,6928	1,6982	1,7037	1,7092
0,82	1,7148	1,7204	1,7260	1,7316	1,7373	1,7430	1,7487	1,7545	1,7603	1,7661
0,83	1,7720	1,7779	1,7838	1,7898	1,7958	1,8018	1,8079	1,8140	1,8202	1,8264
0,84	1,8326	1,8389	1,8452	1,8515	1,8579	1,8643	1,8708	1,8773	1,8839	1,8905
0,85	1,8971	1,9038	1,9105	1,9173	1,9241	1,9310	1,9379	1,9449	1,9519	1,9590
0,86	1,9661	1,9733	1,9805	1,9878	1,9951	2,0025	2,0099	2,0174	2,0250	2,0326
0,87	2,0402	2,0479	2,0557	2,0636	2,0715	2,0795	2,0876	2,0967	2,1038	2,1120
0,88	2,1203	2,1287	2,1371	2,1456	2,1542	2,1629	2,1716	2,1804	2,1893	2,1983
0,89	2,2073	2,2264	2,2256	2,2349	2,2443	2,2538	2,2634	2,2731	2,2829	2,2927
0,90	2,3026	2,3126	2,3228	2,3330	2,3434	2,3539	2,3645	2,3752	2,3860	2,3969
0,91	2,4079	2,4191	2,4304	2,4418	2,4534	2,4651	2,4769	2,4889	2,5010	2,5133
0,92	2,5257	2,5383	2,5510	2,5639	2,5770	2,5903	2,6037	2,6173	2,6311	2,6451
0,93	2,6593	2,6736	2,6882	2,7030	2,7181	2,7334	2,7489	2,7646	2,7806	2,7969
0,94	2,8134	2,8302	2,8473	2,8647	2,8824	2,9004	2,9188	2,9375	2,9565	2,9759
0,95	2,9957	3,0159	3,0366	3,0576	3,0791	3,1011	3,1236	3,1466	3,1701	3,1942
0,96	3,2188	3,2442	3,2702	3,2968	3,3242	3,3524	3,3814	3,4112	3,4420	3,4738
0,97	3,5066	3,5405	3,5756	3,6119	3,6497	3,6889	3,7297	3,7723	3,8167	3,8632
0,98	3,9120	3,9633	4,0174	4,0745	4,1352	4,1997	4,2687	4,3428	4,4228	4,5099
0,99	4,6052	4,7105	4,8283	4,9618	5,1160	5,2983	5,5215	5,8091	6,2146	6,9078
0,999	6,9078	7,0131	7,1309	7,2644	7,4186	7,6009	7,8240	8,1117	8,5172	9,2103
0,9999	9,2103	9,3157	9,4335	9,5670	9,7212	9,9035	10,1266	10,4143	10,8198	11,5129
0,99999	11,5129	11,6183	11,7361	11,8696	12,0238	12,2061	12,4292	12,7169	13,1224	13,8155
0,999999	13,8155	13,9209	14,0387	14,1722	14,3263	14,5087	14,7318	15,0195	15,4249	16,1181

TOMMASO SALVEMINI

Su la determinazione dei quozienti di eliminazione [1]

1. — Premessa. — I quozienti di eliminazione servono tutte le volte che si vuol studiare quantitativamente l'eliminazione prodotta da una causa in una collettività di individui, ciascuno dei quali è soggetto al rischio di essere eliminato per quella causa.

Il caso più comune è che la collettività considerata sia la popolazione censita in uno Stato oppure gli assicurati di una Compagnia, e che la causa di eliminazione sia la morte, ma può darsi benissimo che la collettività comprenda soltanto gli individui di un certo stato civile, o di una data categoria professionale, o gli studenti di una università, o gli appartenenti a una religione, o le donne esposte al rischio di concepimento, o le coppie dei coniugi, ecc. e che l'evento che determina l'eliminazione sia per es. il cambiamento di stato civile, o l'invalidità, o l'emigrazione, o l'abbandono di quel credo religioso, o l'avvenuta fecondazione, o la scissione della coppia, ecc. La causa di eliminazione può essere definitiva o temporanea, singola o data dal concorso di parecchie cause singole. Essendo molteplici e importanti i campi di applicazione, e le collettività statistiche non essendo quasi mai formate da individui omogenei — in quanto differiscono generalmente per l'età o per altri caratteri collegati all'evento di eliminazione — oppure avvenendo in seno ad esse movimenti di entrata e di uscita per cause estranee a quella di eliminazione, la trattazione presenta particolari difficoltà, ed è chiaro, quindi, che l'esatta determinazione dei quozienti di eliminazione — saggio (o tasso) e probabilità di eliminazione — abbia richiamato l'attenzione di personalità scientifiche di primo piano, al fine di dare metodi ap-

[1] Comunicazione presentata il 7 maggio 1946, alla riunione del Seminario di Statistica dell'Istituto di Statistica della Università di Roma.

propriati per il loro calcolo in base ai dati statistici disponibili sul movimento della collettività [1].

Ancora oggi ci sono dibattiti scientifici su questo argomento per cercare di selezionare i metodi esistenti, mettendone in evidenza i pregi e i difetti.

È del 1941 un lavoro critico del Cantelli [2] al quale ha fatto seguito prima uno studio di Del Chiaro [3] e poi uno studio dell'Ottaviani [4], uno dei migliori allievi del Cantelli. L'esame di questi lavori mi ha fatto sorgere il desiderio di un maggiore approfondimento dell'importante questione, ai fini di portare il mio modesto contributo sia a chiarimento o precisazione dei punti controversi, in base ad ulteriore analisi delle ipotesi e dei procedimenti che sono a base dei metodi noti, sia per indicare possibili sviluppi di tali metodi per il calcolo dei quozienti di eliminazione nei casi concreti.

2. — Saggio medio di eliminazione e ipotesi per il calcolo delle probabilità di eliminazione. — Sia C la collettività di individui — o, più in generale, di unità statistiche — in considerazione ; c la causa di eliminazione ; l_t il numero di appartenenti alla collettività al tempo t ; $k_u\,du$ la differenza (positiva o negativa) tra il numero di coloro che entrano in C e il numero di coloro che ne escono per cause diverse da c tra il tempo u e $u + du$; $m_u\,du$ il numero di eliminati dalla collettività per la causa c tra il tempo u e $u + du$; q_t la probabilità pura di eliminazione, ottenuta ragguagliando il numero m_t degli individui

---

[1] In Italia ricordiamo : F. P. Cantelli, *Genesi e costruzione delle tavole di mutualità*, « Bollettino di notizie sul Credito o sulla Previdenza », 1914, n. 3 ; C. Gini, *Su la determinazione dei quozienti di eliminazione e in particolare sui metodi delle durate esatte e delle durate medie nell'ipotesi di saggi istantanei di eliminazione costanti*, « Metron », Vol. XII, n. 3, 1935.

[2] F. P. Cantelli, *Sulla costruzione delle tavole di mortalità*, « Atti della III riunione della Società italiana di statistica », giugno-luglio, 1941.

[3] A. Del Chiaro, *Sui tassi centrali di mortalità*, « Giornale dell'Istituto Italiano degli Attuari », n. 3-4, 1941.

[4] G. Ottaviani, *Sulle tavole di mortalità*, « Giornale dell'Istituto Italiano degli Attuari », n. 1-4, 1942.

eliminati durante l'intervallo $t \vdash t+1$ al numero E_t degli individui eliminabili durante detto intervallo per la causa c ; S_t il saggio di eliminazione, ottenuto ragguagliando il numero m_t degli eliminati alla media M_t degli esposti al rischio durante l'intervallo $t \vdash t+1$; finalmente, μ_u il saggio istantaneo di eliminazione, uguale al rapporto tra $m_u\,du$ e il numero $E_u\,du$; preciseremo meglio nel seguito il significato di μ_u [1].

Premesse queste notazioni, è chiaro che la media degli esposti al rischio durante l'intervallo $t \vdash t+1$ è data da

$$M_t = \int_t^{t+1} E_s\,d\,s \tag{1}$$

e tenendo presente che

$$E_s = l_t + \int_t^s (k_u - m_u)\,du,$$

si avrà, con una opportuna integrazione per parti,

$$M_t = l_t + \int_t^{t+1}\left[\int_t^s (k_u - m_u)\,du\right] ds = l_t + \left[s\int_t^s (k_u - m_u)\,du\right]_t^{t+1} -$$

$$- \int_t^{t+1} s\,(k_s - m_s)\,ds = l_t + (t+1)\int_t^{t+1} (k_u - m_u)\,du -$$

$$- \int_t^{t+1} u\,(k_u - m_u)\,du = l_t + \int_t^{t+1} (k_u - m_u)\,(t+1-u)\,du, \tag{1'}$$

espressione che si poteva ottenere più rapidamente — anche se meno immediatamente per chi ha minore dimestichezza con gli integrali doppi — osservando che

---

[1] Quasi tutti gli attuari usano la denominazione di *tasso* di eliminazione. Riteniamo, invece, di forma più italiana la denominazione di *saggio* di eliminazione usata dal Gini.

$$\int_t^{t+1}\int_t^s (k_u - m_u)\, du\, ds = \int_t^{t+1} (k_u - m_u)\, du \int_u^{t+1} ds =$$

$$= \int_t^{t+1} (k_u - m_u)\, (t + 1 - u)\, du$$

Dalla (1') risulta immediatamente che M_t si ottiene sommando le durate effettive di esposizione al rischio sia degli esistenti al tempo t, sia degli entrati, degli usciti e degli eliminati da C tra t e $t + 1$.

Pertanto, il saggio medio di eliminazione S_t sarà dato da:

$$S_t = \frac{m_t}{l_t + \int_t^{t+1} (k_u - m_u)\, (t + 1 - u)\, du} \qquad (2)$$

Questa espressione, evidentemente, non è subordinata ad alcuna ipotesi.

Nei casi concreti, però, se non si conosce il momento preciso di entrata e di uscita per cause diverse da c, e quello di eliminazione per la causa c di ciascun individuo della collettività, allora, per il calcolo di S_t, o si ricorre a formule approssimate e convenienti di quadratura dell'integrale che figura nella (2), sempre che si abbiano gli elementi per un tale procedimento, oppure si ricorre a introdurre un'ipotesi su questo movimento di entrata, di uscita e di eliminazione, la quale permetta il calcolo di detto integrale.

Osserviamo, intanto, che il denominatore della (2) può essere anche notevolmente inferiore al numeratore m_t per cui il saggio medio S_t può essere maggiore di uno. Se ad es. in una collettività di 100 individui si ha una uniforme eliminazione di tutti questi individui nel corso di un periodo, allora $k_u = 0$, m_u è costante e quindi $S_t = \frac{100}{100 - 50} = 2$, mentre la probabilità pura di eliminazione è evidentemente uguale all'unità.

Ne segue che il confronto tra saggi medi che si riferiscono a collettività diverse (nello spazio o nel tempo) ha gli stessi in-

convenienti dei confronti tra indici assoluti ; determinare il massimo di S_t (il quale si ha quando gli eliminati m_t nel periodo sono in numero uguale agli esposti E_t) e rapportare S_t a questo massimo equivale a calcolare la probabilità di eliminazione q_t.

Ora, per ottenere questa probabilità di eliminazione, se il gruppo C è chiuso (se cioè in essa non ci sono movimenti di entrata e di uscita per cause diverse da quella di eliminazione, in ogni intervallo) e se tutti gli individui della collettività sono egualmente eliminabili, allora $E_t = l_t$ e quindi

$$q_t = \frac{m_t}{l_t}$$

Queste circostanze si verificano assai raramente. In particolare se la collettività non è chiusa, non si potranno considerare egualmente eliminabili due individui che abbiano durata diversa di esposizione al rischio. D'altra parte, per poter stabilire con esattezza la quota di rischio di ciascuno di detti individui occorre introdurre una qualche ipotesi sulla legge di eliminazione incognita che si cerca di studiare proprio mediante i quozienti detti. Se per di più non si conosce il momento preciso di entrata e di uscita di ciascuno dalla collettività per cause diverse da c, occorre anche ricorrere ai procedimenti indicati precedentemente per il calcolo di S_t.

Questo è un grave inconveniente, per modo che è essenziale conoscere bene le ipotesi che sono a base dei metodi per il calcolo delle probabilità q_t, per evitare errate conclusioni che possono trarsi dai confronti fra queste probabilità. Infatti per collettività diverse a confronto possono non sussistere le stesse condizioni che rendono consigliabile far ricorso alle stesse ipotesi ; bisogna inoltre badare che le differenze tra probabilità a confronto siano proprio significative e non dipendano invece dalle ipotesi che sono a base del loro calcolo [1].

---

[1] Facciamo astrazione dalle deviazioni casuali, di cui sono affetti i q_t, dovute, di solito, al numero ristretto dei casi osservati. È compito della perequazione di compensare detta deviazione ; se ogni collettività in esame è abbastanza ampia e omogenea, le deviazioni casuali si compensano abbastanza bene.

In mancanza di cognizioni precise sulla validità o meno di queste ipotesi è certo da statistici avveduti e prudenti eseguire il calcolo con due o più di esse ad esaminare sino a qual punto certe deduzioni possano essere accettate.

Circa le diverse ipotesi sulla legge di eliminazione, esaminiamo con Gini [1] le seguenti:

a) il numero di eliminati per la causa c sia uguale nei successivi istanti dell'intervallo $t \vdash t + 1$, ciò che equivale — supposto che gli esposti al rischio restino in osservazione durante tutto l'intervallo — a supporre che il saggio istantaneo di eliminazione vada crescendo in progressione armonica dal principio alla fine del periodo di osservazione;

b) nei successivi istanti dell'intervallo $t \vdash t + 1$, rimanga costante non il numero degli eliminati per la causa c ma il saggio istantaneo μ_u di eliminazione;

c) l'eliminazione avvenga nel primo istante;

d) n individui in osservazione incompleta, i quali entrino in osservazione nel momento u dell'intervallo unitario $t \vdash t + 1$ e sono seguiti sino alla fine dell'intervallo, equivalgono, ai fini del calcolo della probabilità di eliminazione, a $(t + 1 - u)\, n$ individui in osservazione completa;

e) il numero degli eliminati varii in funzione lineare del tempo durante l'intervallo $t \vdash t + 1$.

Tra tutte queste ipotesi, la *a*) è spesso usata per determinare il coefficiente di mortalità quando gli esposti al rischio restino in osservazione durante tutto l'intervallo $t \vdash t + 1$, e su essa è basata la nota formula di Becker.

La *d*) fu ammessa dall'Ackland e sta alla base del metodo delle durate esatte, applicato quando nel gruppo osservato si verificano movimenti di entrata e di uscita all'infuori della causa c [2].

In base ad essa si ha:

$$\bar{q}_t = \frac{m_t}{l_t + \int_t^{t+1} k_u\,(t + 1 - u)\, du} \qquad (3)$$

---

[1] C. Gini, *Su la determinazione...*, op. cit.

[2] Vedasi bibliografia nel lavoro cit. del Gini.

Il Cantelli (1), facendo l'analisi di questa formula dal punto di vista probabilistico, mise in evidenza che essa diventa esatta nell'ipotesi che la probabilità di eliminazione $q(u, t+1)$ dall'istante u all'istante $t+1$ sia proporzionale all'intervallo $t+1-u$, cioè sia tale che si abbia

$$q(u, t+1) = q_t(t+1-u) \tag{4}$$

Pertanto la (3) è da ritenersi approssimata non soltanto per il sistema di anno di calendario (come riteneva l'Ackland stesso) ma per tutti i casi che non soddisfano la (4), Da quest'ultima relazione si deduce, poi, che la (3) viene giustificata soltanto per un saggio di eliminazione decrescente e ciò — come ha osservato il Cantelli e ha riconfermato il Gini — è per lo più in contrasto con la realtà quando si tratti di determinare la probabilità di eliminazione per morte, in quanto la mortalità cresce col crescere dell'età dopo i primi anni di vita, e, per molte età di entrata negli Istituti di assicurazione, cresce, generalmente, con la durata dell'assicurazione; gli attuari, però, considerano la (3) come sufficientemente approssimata per i loro scopi pratici; per cause di eliminazione diverse da quella di morte, essa potrebbe indurre a sensibili errori.

L'ipotesi *b*) è quella su cui si è particolarmente fermato il Gini nel lav. cit. Essa pur non pretendendo di essere sempre applicabile, spesso è quella che meglio si avvicina alla realtà, ed appare giustificata, in base al principio di indifferenza, quando non si conosca nulla circa la variazione di detto saggio durante l'intervallo considerato.

In base a questa ipotesi, dalla nota relazione:

$$q_t = 1 - e^{-\int_t^{t+1} \mu_u \, du} \tag{5}$$

e ponendo $\mu_u = S_t = \text{cost.}$ in ogni intervallo, o con ragionamento diretto (2), si ha:

$$q_t = 1 - e^{-S_t} \tag{6}$$

---

(1) **F. P. Cantelli**, *Genesi e costruzione...*, op. cit.

(2) C. Gini, *loc. cit.*, in (1), pagg. 5-7.

L'ipotesi *b*) era stata già considerata dal Wittstein [1] il quale se ne serviva per determinare gli esposti al rischio con un procedimento approssimato non rispondente in pratica.

Diverso è quindi il metodo qui indicato che può chiamarsi *metodo delle durate esatte del Gini*, di cui lo stesso autore ha determinato anche casi particolari molto interessanti, oltre a farne importanti applicazioni.

3. — ALCUNE RISERVE SUL METODO DELLE DURATE ESATTE DEL GINI. — Il Cantelli, partendo da altre considerazione — che per brevità non sto a ripetere — dimostra, supponendo μ_s variabile, la formula

$$\int_t^{t+1} \mu_u \, du = \frac{m_t}{l_t + \int_t^{t+1} (k_u - m_u) f_1(u, t+1) \, du} \tag{7}$$

dove $f_1(u, t+1)$ è una funzione positiva decrescente tale che:

$$\int_u^{t+1} \mu_s \, d\, s = \int_t^{t+1} \mu_s \, d\, s \cdot f_1(u, t+1) \tag{8}$$

Dalla (5) e dalla (7) si ricava la probabilità di eliminazione.

Vediamo ora le critiche del Cantelli e dell'Ottaviani.

Essi giustamente osservano che se la $f_1(u, t+1)$, tenuta presente la (8), non fornisce una buona approssimazione della funzione $\int_t^{t+1} \mu_s \, ds$, la (5) con la (7) conduce a valori erronei della probabilità di eliminazione. La cattiva approssimazione della f_1 si riflette non solo sul movimento di entrata e uscita degli individui dalla collettività C, ma anche sugli eliminati dal gruppo iniziale per la causa c. Nella (3), invece, l'approssimazione della $t + 1 - u$ (che dovrebbe essere invece sostituita — come mostra il Cantelli — da una funzione $f(u, t) = \frac{q(u, t+1)}{q_t}$, decre-

---

[1] T. WITTSTEIN, *Die Mortalität in Gesellschaften mit successiv eintretenden und ausschreidenden Mitgliedern*, in «Grunert's Archiv der Mathematik und Physik» Bd 39, 1862.

scente da 1 a zero al crescere di u da t a $t + 1$) si riflette solamente sul movimento degli ingressi e delle eliminazione per cause diverse dalla c ; quindi se k_u assume valori piccoli (o nulli) la (3) fornisce una notevole approssimazione (o il valore esatto) della probabilità di eliminazione.

Da ciò la preferenza che essi danno alla (3) in confronto alla (5) con la (7), e così pure in confronto all'ipotesi $\mu = \text{cost.} = S_t$, nel qual caso, dalla (8),

$$f_1(u, t+1) = \frac{\int_u^{t+1} \mu_s \, ds}{\int_t^{t+1} \mu_s \, ds} = t + 1 - u \qquad (8')$$

Analizziamo ora queste osservazioni nel caso delle formule che interessano il metodo indicato dal Gini.

4. — Relazione tra la (2) e la (7). — Intanto osserviamo che il primo membro della (2) coincide col primo membro della (7) quando si fa l'ipotesi che il saggio istantaneo di eliminazione resti costante durante tutto l'intervallo. In tal caso, da quanto si è visto, la $f_1(u, t + 1)$ *deve essere proprio uguale a* $t + 1 - u$; viceversa se $f_1(u, t + 1) = t + 1 - u$ il saggio istantaneo di eliminazione è indipendente da u, cioè costante nell'intervallo $t \vdash t + 1$ [1].

Pertanto la (8') non è una condizione restrittiva che rende meno esatto il calcolo dell'integrale che figura nella (2), in quanto in essa non può sostituirsi $t + 1 - u$ con un'altra funzione. La (8') è semplicemente una condizione equivalente a quella del Gini.

Ma per meglio chiarire la differenza tra la (2) e la (7) osserviamo ancora che la (7) — nella quale la f_1 sia data dalla (8) e tenuto conto che l'intervallo è unitario — rappresenta *la media aritmetica semplice* dei saggi istantanei di eliminazione. Se invece teniamo presente che :

$$m_s \, d\, s = \mu_s \, E_s \, ds \qquad (9)$$

---

[1] F. P. Cantelli, *Sulla costruzione* ...., cit. p. 23.

e facciamo nell'intervallo $t \vdash t+1$ *la media aritmetica ponderata* dei μ_s con pesi uguali ad E_s, si ha la quantità

$$\frac{\int_t^{t+1} \mu_s E_s \, ds}{\int_t^{t+1} E_s \, ds} = \frac{\int_t^{t+1} m_s \, ds}{\int_t^{t+1} E_s \, ds} = \frac{m_t}{\int_t^{t+1} E_s \, ds} = S_t \qquad (9')$$

La (2) è dunque sostanzialmente diversa dalla (7) in quanto mentre la prima rappresenta una media ponderata dei saggi istantanei di eliminazione, la seconda espressione rappresenta una media semplice. Le due medie diventano uguali se μ_s è costante se cioè $f_1(u, t+1) = t + 1 - u$. Se quindi nella (2) poniamo al posto di $t + 1 - u$ la f_1 data dalla (8), essa perde il significato di media aritmetica ponderata dei μ_s per diventare una media aritmetica semplice di tali saggi istantanei; ed è chiaro, in tal caso, la superiorità della media ponderata in confronto della media semplice per dare una misura media dell'effetto della causa di eliminazione in tutto l'intervallo $t \vdash t+1$.

Se, inoltre, a causa della quasi impossibilità di conoscere esattamente, nei casi concreti, l'espressione della $f_1(u, t+1)$, si sostituisce ad essa una sua espressione approssimata diversa da $t + 1 - u$, allora, dalla (7) non si otterebbe più nemmeno il valore esatto della media aritmetica dei saggi di eliminazione, mentre la (2) dà, di per sè, indipendentemente da alcuna condizione restrittiva, un elemento importante di conoscenza per lo studio della causa di eliminazione.

Aggiungiamo che se teniamo presente il significato di M_t esposto nel n. 2, si ha chiaramente che $\bar{q}_t$, dato dall'Ackland mediante la (3), non è che un saggio medio di eliminazione nella ipotesi che ciascun eliminato per la causa c non esca dalla collettività all'istante dell'eliminazione ma al tempo $t + 1$; ipotesi, quest'ultima, poco aderente alla realtà e quindi molto approssimata. Questo spiega il fatto che $\bar{q}_t$ può essere alle volte, maggiore di 1, come di fatto si è verificato nelle applicazioni del Gini [1].

---

[1] C. Gini, *Su la determinazione...*, op. cit.

Se questa stessa quantità (3) la si vuole legittimare come probabilità di eliminazione, allora è necessaria, come si è già detto (1), la condizione che la probabilità di morte tra l'età u e $t+1$ (con $t \leqslant u < t+1$) sia data dalla espressione $q_t\,(t+1-u)$.

Il saggio medio di eliminazione, in un intervallo unitario di tempo, è paragonabile alla velocità media in un moto vario la cui velocità all'istante (saggio istantaneo nel nostro caso) varia da istante a istante. Se il rapporto verificatosi all'istante u tra eliminati ed esposti al rischio rimanesse costante per tutti gli istanti dell'intervallo, il saggio medio sarebbe uguale al saggio istantaneo e viceversa. Inoltre dalla (9') risulta che S_t è una media interna dei μ_s (è quindi compreso tra il minimo e il massimo valore di μ_s) e poichè, come abbiamo già messo in evidenza a proposito della (2), S_t può essere maggiore dell'unità, ne segue che anche μ_s può essere maggiore dell'unità e ciò riconferma che questa quantità non è una probabilità ; mentre è tale la quantità $\mu_s\,ds$ a meno di infinitesimi di ordine superiore a ds.

In effetti μ_s è definito o dalla (9) o dalla relazione equivalente :

$$\mu_s = -\frac{d}{ds}\log E(s) = \lim_{\Delta s \to 0} \frac{E_s - E_{s+\Delta s}}{E_s\,\Delta s} \qquad (9'')$$

supposta la funzione $E(s)$ continua e derivabile della variabile s, e non crescente al crescere di s a causa delle eliminazioni. Ora, mentre la quantità $\frac{E_s - E_{s+\Delta s}}{E_s}$ non può essere maggiore di uno, il suo rapporto a Δs può essere certamente > 1. Ciò capita in particolare quando la causa di eliminazione è molto intensa all'inizio s del periodo, per modo che la derivata prima della funzione $E(s)$ nel punto s sia maggiore dell'ordinata E_s.

Il fatto che μ_s è uguale al rapporto tra una probabilità e Δs conferma che il saggio istantaneo di eliminazione nell'istante s fornisce, in realtà, «il valore del saggio di eliminazione che si otterrebbe in un intervallo *unitario*, qualora persistesse, per tutti gli istanti dell'intervallo, il rapporto tra eliminati e esposti al rischio verificatosi nell'istante s» (2).

---

(1) F. P. Cantelli, *Genesi e costruzione...*, cit. pag. 37.

(2) C. Gini, *loc. cit.* pag. 4, **nota.**

5. — Sul calcolo di $\int_{t}^{t+1}(k_u - m_u)\,(t + 1 - u)\,du$. — Nella pratica, il calcolo numerico dell'integrale che figura nella (2) è generalmente faticoso in quanto non si conosce l'espressione analitica esatta di $k_u - m_u$, per cui — se non si conoscono o non si vuol ricorrere alle durate esatte di esposizione al rischio dei singoli individui della collettività e se non si vuol ricorrere ad ipotesi supplementari sull'andamento di k_u e di m_u, o quanto meno di $k_u - m_u$ — bisogna rivolgersi, come si è detto, a formule di integrazione approssimata in base alla conoscenza statistica del movimento di entrate e di uscita da C per cause diverse da quella di eliminazione, e del movimento di uscita per la causa c, in corrispondenza ad intervalli parziali contenuti in $t \vdash t + 1$.

Al riguardo, è noto che la matematica offre una larga gamma di formule di quadrature da scegliere appropriatamente all'andamento reale o plausibile della funzione integranda. Che questa sia k_u soltanto o sia $k_u - m_u$, l'essenziale è scegliere una formula appropriata di quadratura, compatibile anche con i dati statistici di cui si può disporre.

Se quindi, per la formula (3), in relazione ai dati disponibili o al reale andamento di k_u non c'è una formula di quadratura che vada bene, i risultati che da essa si ottengono saranno poco esatti, altrettanto come avverrebbe se non è consentita una buona approssimazione dell'integrale che figura nella (2).

Nè si può dire che l'andamento analitico di k_u sia da ritenere più semplice di quello di $k_u - m_u$, per modo che sia più facile avere una buona approssimazione per

$$\int_{t}^{t+1} k_u\,(t + 1 - u)\,du \quad \text{piuttosto che per} \quad \int_{t}^{t+1}(k_u - m_u)\,(t + 1 - u)\,du \qquad (a)$$

L'Ottaviani, ad esempio, ha indicato altre formule teoriche per la determinazione di q_t, ma esse, come egli stesso osserva, sono subordinate all'integrazione approssimata. Aggiungiamo che sulle funzioni integrande che figurano in queste formule, è in generale più difficile fare ipotesi ragionevoli e semplici, su cui facilmente convenire, per poter scegliere la formula di quadratura più approssimata e per poter valutare in qualche modo il

resto — da cui l'errore per q_t — che si ottiene usando tali formule approssimate di integrazione.

Con ciò non vogliamo escludere che se k_u assume valori molto piccoli l'approssimazione nel primo integrale della (*a*) sia generalmente migliore di quella del secondo, ma il contrario avviene se, indipendentemente dall'altezza di k_u si abbia che $k_u - m_u$ assume valori molto piccoli. Confrontando ad es. la tav. XIV del lav. cit. (1) del Gini o la nostra tav. II al n. 13 di questo lavoro, si vede che per la causa di eliminazione ivi studiata, i valori di $k_t - m_t$ sono molto più piccoli dei k_t, ad eccezione delle ultime età, e quindi avviene proprio che la funzione integranda che figura nella (3) è più grande di quella che figura nella (2).

Possiamo però ancora spingere più in là le nostre osservazioni al riguardo, se teniamo presente che, contrariamente alla (3), per avere q_t in base al valore di S_t calcolato con la (2) bisogna poi servirsi della (6).

Indichiamo con α l'errore che si ottiene per la probabilità di eliminazione usando una determinata formula di quadratura per il calcolo dell'integrale che figura nella (3) ; sia cioè

$$\bar{q}_t' = \bar{q}_t + \alpha \qquad (10)$$

dove $\bar{q}'_t$ indica il valore approssimato ottenuto invece di $\bar{q}_t$.

Supponiamo poi che con la stessa formula di quadratura o con altra formula più approssimata per l'integrale che figura nella (2) si commetta lo stesso errore α nel calcolo di S_t, sia cioè :

$$S'_t = S_t + \alpha \qquad (10')$$

In conseguenza di questo valore approssimato di S_t avremo per la (6)

$$q'_t = 1 - e^{-(St+\alpha)} = 1 - e^{-St}\, e^{-\alpha}$$

e quindi :

$$q'_t - q_t = e^{-St}\,(1 - e^{-\alpha}) \qquad (10'')$$

dove, ora, questa espressione rappresenta l'errore che si commette su q_t, adoperando il metodo del Gini, nella supposizione che α sia l'errore ottenuto tanto per $\bar{q}_t$ quanto per S_t a causa della formula approssimata di integrazione.

Confrontiamo questi due errori. Essi hanno certamente lo stesso segno perchè — a parte il fattore e^{-S_t}, sempre positivo nella

(10'') — se $\alpha > 0$ si ha $e^{-\alpha} < 1$ e quindi $1 - e^{-\alpha} > 0$, mentre se $\alpha < 0$ si ha $e^{-\alpha} > 1$ e quindi $1 - e^{-\alpha} < 0$. Si ha poi

$$1 - e^{-\alpha} < \alpha \qquad (11)$$

in quanto essa equivale a :

$$e^{-\alpha} > 1 - \alpha$$

che è sempre vera poichè $1 - \alpha$ rappresenta lo sviluppo in serie di $e^{-\alpha}$ arrestato ai primi due termini, e che l'errore commesso, arrestandoci ad essi, è positivo.

Se α è positivo, $1 - e^{-\alpha}$ è positivo e quindi, moltiplicando questa quantità per e^{-S_t} che è certamente minore dell'unità e tanto più piccolo quanto più cresce S_t, risulta in modo più accentuato la più piccola entità dell'errore (10'') in confronto all'errore α ammesso per S_t e per $\bar{q}_t$.

Se, invece, α è negativo, allora $1 - e^{-\alpha}$ è negativo e la diseguaglianza (11) si verifica soltanto in valore algebrico ; quando poi moltiplichiamo il primo membro per la quantità , più piccola dell'unità, e^{-S_t}, il segno della (11) può invertirsi, nel qual caso l'errore di q_t torna ad essere inferiore, in valore assoluto, a quello di $\bar{q}_t$.

Sia nel primo che nel secondo caso, per facilitare il confronto, in valore algebrico e in valore assoluto, tra l'errore α e l'errore (10''), e per vedere l'entità delle differenze tra essi, tabelliamo il valore (10'') per alcuni valori di S_t e di α.

Ci limitiamo a considerare un errore α compreso tra — 0,5 e + 0,5 perchè è chiaro che valori di α più elevati in valore assoluto di 0,5 corrispondono ad approssimazioni da escludere senz'altro.

Da questa tabella si vede subito che, al crescere di S_t, l'errore (10''), preso in valore assoluto, va continuamente decrescendo ; pertanto, in corrispondenza a valori elevati della probabilità di eliminazione, un errore α — sia pure elevato — di S_t non ha sensibile influenza su tale probabilità. Il valore assoluto di detto errore (10'') è sempre inferiore ad α quando $\alpha > 0$, mentre avviene generalmente il contrario quando α è negativo e in valore assoluto maggiore di S_t, caso, questo, da escludere perchè, per la (10'), si avrebbe S'_t negativo ; (nella tabella precedente sarebbero perciò da escludere i numeri al di sotto della spezzata a gradinata).

Valori della funzione $e^{-S_t}(1-e^{-\alpha})$ Tav. I.

α \ S_t	0,01	0,05	0,1	0,2	0,3	0,4	0,5	1,0	1,5	2,0
0,5	0,3891	0,3737	0,3557	0,3210	0,2912	0,2633	0,2386	0,1446	0,0876	0,0531
0,4	0,3267	0,3138	0,2987	0,2703	0,2445	0,2211	0,2003	0,1214	0,0736	0,0446
0,3	0,2564	0,2463	0,2344	0,2121	0,1919	0,1735	0,1572	0,0953	0,0578	0,0350
0,2	0,1792	0,1721	0,1638	0,1482	0,1341	0,1213	0,1099	0,0666	0,0404	0,0244
0,1	0,0941	0,0903	0,0860	0,0778	0,0704	0,0637	0,0577	0,0350	0,0212	0,0128
0,05	0,0485	0,0466	0,0443	0,0401	0,0363	0,0328	0,0297	0,0180	0,0109	0,0066
0,02	0,0198	0,0190	0,0181	0,0164	0,0148	0,0134	0,0121	0,0074	0,0045	0,0027
0,01	0,0099	0,0095	0,0091	0,0082	0,0074	0,0067	0,0061	0,0037	0,0022	0,0014
0,005	0,0050	0,0048	0,0045	0,0041	0,0037	0,0034	0,0030	0,0018	0,0011	0,0007
— 0,005	— 0,0050	— 0,0048	— 0,0045	— 0,0041	— 0,0037	— 0,0034	— 0,0030	— 0,0018	— 0,0011	— 0,0007
— 0,01	— 0,0099	— 0,0095	— 0,0091	— 0,0082	— 0,0074	— 0,0067	— 0,0061	— 0,0037	— 0,0022	— 0,0014
— 0,02	— 0,0198	— 0,0190	— 0,0181	— 0,0164	— 0,0148	— 0,0134	— 0,0121	— 0,0074	— 0,0045	— 0,0027
— 0,05	— 0,0505	— 0,0485	— 0,0462	— 0,0418	— 0,0378	— 0,0342	— 0,0310	— 0,0188	— 0,0114	— 0,0069
— 0,1	— 0,1040	— 0,0999	— 0,0950	— 0,0860	— 0,0778	— 0,0704	— 0,0637	— 0,0386	— 0,0234	— 0,0142
— 0,2	— 0,2188	— 0,2102	— 0,2000	— 0,1810	— 0,1638	— 0,1481	— 0,1341	— 0,0813	— 0,0493	— 0,0298
— 0,3	— 0,3465	— 0,3329	— 0,3168	— 0,2867	— 0,2594	— 0,2345	— 0,2125	— 0,1288	— 0,0781	— 0,0473
— 0,4	— 0,4871	— 0,4679	— 0,4453	— 0,4029	— 0,3646	— 0,3296	— 0,2986	— 0,1811	— 0,1097	— 0,0664
— 0,5	— 0,6425	— 0,6172	— 0,5873	— 0,5315	— 0,4809	— 0,4348	— 0,3939	— 0,2388	— 0,1447	— 0,0876

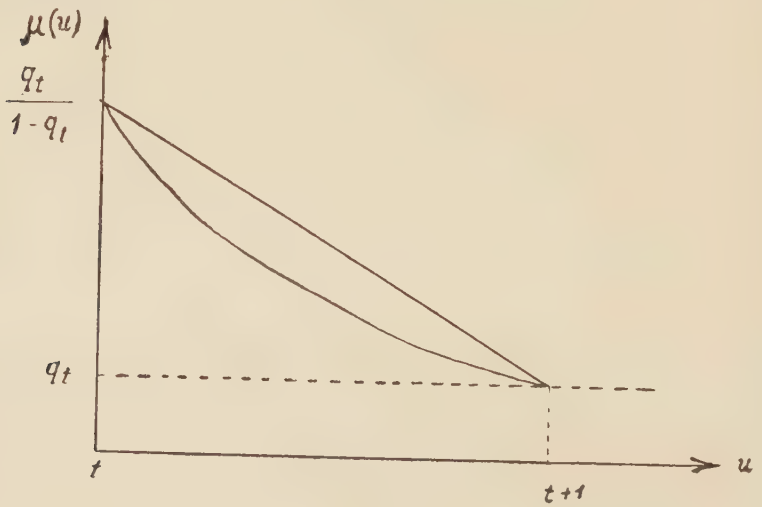
Possiamo quindi concludere, che a parità di errore α per la (2) e la (3) — causa dell'uso di formule approssimate di integrazione per il loro calcolo — l'errore (10″) per q_t dedotto nell'ipotesi di un saggio istantaneo di eliminazione costante è, in valore assoluto, sempre inferiore — o tutt'al più uguale per valori molto bassi di $|\alpha|$ e di S_t — all'errore $|\alpha|$ di $\bar{q}_t$ ricavato col metodo delle durate esatte.

6. — La diminuzione che subisce $\mu(u)$ in ogni intervallo unitario nell'ipotesi $q(u, t+1) = q_t(t+1-u)$. — Riteniamo opportuno spendere ancora qualche parola sulla complessa analisi per il calcolo delle probabilità di eliminazione.

Ricordiamo che nell'ipotesi (4) suddetta, il saggio istantaneo $\mu(u)$ nell'intervallo $t \vdash t+1$ è dato da (1).

$$\mu(u) = \frac{q_t}{1 - q_t(t+1-u)} \qquad (12)$$

La derivata prima di questa funzione è negativa, mentre la derivata seconda è positiva, per cui nell'intervallo $t \vdash t+1$, detta funzione sta sempre al disotto della sua corda, cioè essa ha un andamento rappresentato schematicamente dalla linea curva nella seguente figura:



(1) F. P. Cantelli, *Genesi e costruzione...*, cit.

È da rilevare che la differenza tra $\frac{q_t}{1 - q_t}$ e q_t, cioè la diminuzione che subisce $\mu(u)$ quando u passa da t a $t + 1$, è tutt'altro che trascurabile — come può vedersi dalla seguente tabellina — quando q_t ha valori elevati, come spesso avviene per cause di eliminazione diverse da quella di morte:

q_t	$\frac{1-q_t}{q_t}$
0,01	0,0101
0,02	0,0204
0,05	0,0526
0,10	0,1111
0,20	0,2500
0,50	1,0000
0,70	2,3333
0,90	9,0000

In tali circostanze è quindi evidente che non sia punto accettabile l'ipotesi (4); tanto più quando si abbia ragione di ritenere che il saggio non debba decrescere nei singoli intervalli a cui si riferisce la probabilità.

Se invece facciamo l'ipotesi del Gini di un saggio istantaneo $\mu = \text{costante} = S_t$ (dato dalla (2)), tra t e $t + 1$, si ha:

$$f(u) \cdot \frac{q(u, t+1)}{q_t} = \frac{1 - e^{-\int_u^{t+1} \mu \, ds}}{1 - e^{-\int_t^{t+1} \mu \, ds}} = \frac{1 - e^{-St\,(t+1-u)}}{1 - e^{-St}} \qquad (13)$$

che, per $t \leq u < t + 1$, è una funzione decrescente da 1 a 0, come la $t + 1 - u$ nella (4), ma si mantiene sempre al disopra di questa funzione lineare, che rappresenta la sua corda, come risulta subito calcolando la derivata prima e la derivata seconda

di $f(u)$ ed osservando che queste sono sempre negative per S_t positivo. La $f(u)$ si discosta poco dall'andamento lineare per valori piccoli di S_t, allontanandosene al crescere di S_t come risulta immediatamente calcolando $f(t + \frac{1}{2})$ in corrispondenza ai valori di S_t qui sotto indicati, e osservando che i risultati si allontanano sempre più da 0,5.

S_t	$f(t+\frac{1}{2})$
0,02	0,5050
0,06	0,5086
0,10	0,5126
0,20	0,5251
0,50	0,5621
1,00	0,6225
2,00	0,7310

Ma quando non si hanno più precise conoscenze sulla legge di eliminazione, non possiamo dire se per la $f(u)$ sia più esatto l'andamento lineare o quello rappresentato dalla curva della precedente funzione. Essenziale è però notare che questi andamenti sono entrambi accettabili, cioè entrambi realistici perchè soddisfano alla condizione di decrescenza della $f(u)$ da 1 a 0. Nel caso precedente, invece, avevamo ottenuto per $\mu(u)$ un andamento non accettabile.

Mi pare quindi chiaro che nei frequenti casi in cui l'ipotesi che sta a base del metodo delle durate esatte sia inaccettabile, il metodo del Gini — quand'anche non si sappia se la sua ipotesi è conforme alla legge di eliminazione — non porta certamente a risultati (sia pure approssimati) a priori inaccettabili.

Aggiungiamo inoltre che l'ipotesi del Gini equivale evidentemente ad approssimare la curva rappresentatrice del saggio istantaneo mediante una spezzata a gradinata; questo modo di procedere, oltre ad essere comunemente usato in statistica, è usato anche dagli stessi attuari che in pratica attribuiscono la stessa probabilità di morte q_t a tutti gli individui di età compresa tra

t e $t+1$, anche se in effetti la probabilità di morte varia col variare continuo dell'età.

7. — E_t ED M_t NELL'IPOTESI DEL SAGGIO COSTANTE E DI UNIFORME DISTRIBUZIONE DEI k_u. — Sempre nell'ipotesi $\mu = \text{cost.}$ in ciascun intervallo unitario, possiamo scrivere qualche relazione interessante.

È noto che la formula esatta per il calcolo degli esposti al rischio E_t (a cui si deve rapportare m_t per avere q_t) è

$$E_t = l_t + \int_t^{t+1} k_u \frac{q\,(u,\, t+1)}{q_t}\, du \tag{14}$$

Per la (13) si ha:

$$E_t = l_t + \int_t^{t+1} k_u \frac{1 - e^{-St\,(t+1-u)}}{1 - e^{-St}}\, du \tag{15}$$

e se supponiamo, come generalmente si ammette, sia pure con una certa approssimazione, che k_u sia costante, uguale a k_t, nello intervallo $t \vdash t+1$, si ha:

(16)

$$E_t = l_t + \frac{k_t}{1 - e^{-St}} - \frac{k_t}{1 - e^{-St}} \int_t^{t+1} e^{-St(t+1-u)}\, du =$$

$$= l_t + \frac{k_t}{1 - e^{-St}} - \frac{k_t}{S_t}$$

e se indichiamo con q_t la probabilità di eliminazione nell'ipotesi di un saggio costante tra t e $t+1$ si ha:

$$E_t = l_t + k_t \left(\frac{1}{q_t} - \frac{1}{S_t} \right) \tag{16'}$$

nella quale, ricordiamo (1), la differenza $\frac{1}{q_t} - \frac{1}{S_t}$ non supera mai l'unità e tende a $\frac{1}{2}$ col diminuire di S_t.

---

(1) Cfr. C. GINI, *Op. cit.*, pagg. 37-38.

Sotto le stesse ipotesi si ha evidentemente:

$$m_t = q_t E_t = (1 - e^{-St}) \left[l_t + \frac{k_t}{1 - e^{-St}} - \frac{k_t}{S_t} \right] \qquad (17)$$

e inoltre:

$$M_t = \frac{m_t}{S_t} = \frac{(1 - e^{-St})}{S_t} \left(l_t + \frac{k_t}{q_t} - \frac{k_t}{S_t} \right) \qquad (18)$$

8. — Ipotesi semplificatrici sull'andamento di $k_u - m_u$. — Veniamo ora ad un'altra osservazione. Quando non si conoscono i momenti precisi di entrata, di uscita e di eliminazione di ciascun individuo dalla collettività, per l'integrazione che figura nella (2) e nella (3), se non si vuole o non si può ricorrere a formule di quadratura (per la mancanza di elementi statistici relativi a periodi compresi tra t e $t + 1$) occorrerebbe, come abbiamo già detto, conoscere l'espressione analitica di $k_u - m_u$. In mancanza di questa si suole aggiungere un'altra ipotesi alle rispettive ipotesi *b*) o *d*).

Se nella (3) si suppone l'uniforme distribuzione dei k_u, e si pone quindi $k_t = \int_t^{t+1} k_u \, d_u$ — determinabile statisticamente in quanto k_t rappresenta la differenza (positiva o negativa) tra il numero di entrati e il numero di usciti dalla collettività C per cause estranee a c nell'intervallo di tempo $t \vdash t + 1$ — si ha il *metodo delle durate medie.*

In analogia a questo procedimento il Gini suppone nella (2) l'uniforme distribuzione dei $k_u - m_u$ durante l'anno di durata, ed ottiene:

$${}_aS_t = \frac{m_t}{l_t + \frac{k_t - m_t}{2}} \qquad (19)$$

dalla quale, con la (6), si risale poi alla probabilità di eliminazione ${}_aq_t$.

Lo stesso Gini osserva però che detta ipotesi « è, a rigore, contraddittoria coll'ipotesi della costanza dei saggi istantanei di eliminazione che sta alla base del suo metodo, in quanto essa

suppone che nell'intervallo $t \vdash t+1$ si distribuiscono uniformemente, non solo gli entrati e gli usciti, ma anche gli eliminati, mentre, se i saggi istantanei di eliminazione sono costanti, il numero degli eliminati tra gli esposti al rischio all'inizio t dell'intervallo diminuisce nel tempo proporzionalmente al diminuire di tali esposti al rischio, mentre, conviene, d'altra parte, aggiungere gli eliminati provenienti dagli entrati e togliere quelli che sarebbero provenuti dagli usciti.

In realtà l'obiezione sarebbe infondata, in quanto la semplificazione $t+1-u = \frac{1}{2}$ non suppone precisamente che nell'intervallo $t \vdash t+1$ si distribuiscono uniformemente gli entrati, gli usciti e gli eliminati, ma solo che si distribuiscano uniformemente le differenze tra i numeri degli entrati e degli usciti o eliminati, ipotesi che non è contraddittoria con l'altra che il numero degli eliminati varii proporzionalmente al numero degli esposti al rischio ».

Inoltre, la particolare forma lineare che deve avere k_u nell'ipotesi *b*) e della uniforme distribuzione di $k_u - m_u$, ricavata dal Cantelli (*Sulla costruzione*..., cit.) mostra che se tale andamento fosse soddisfatto, allora la (19) darebbe il valore esatto di S_t; in caso contrario la differenza tra S_t ed ${}_aS_t$ dipenderà dal discostarsi di k_u da detto andamento. Nei casi concreti è generalmente impossibile verificare se k_u soddisfa o meno alla sua espressione teorica confacente alle dette ipotesi in tutto l'intervallo $t \vdash t+1$.

D'altra parte, osserva il Gini, l'ipotesi semplificatrice della costanza di $k_u - m_u$ molte volte non risponde a verità e quindi può essere consigliabile non ricorrere ad essa : in taluni casi è più plausibile ammettere che k_u e m_u (o quanto meno la loro differenza) siano proporzionali al numero degli esposti al rischio secondo la legge di eliminazione. Pertanto, in contrapposto al noto metodo delle durate medie c'è il metodo delle durate medie del Gini, fondato sull'ultima ipotesi detta oltre alla *b*), ipotesi analoghe sia per la distribuzione degli eliminati, sia degli entrati ed usciti.

Questa coerenza che dà pregio al metodo non significa — avverte lo stesso Gini — che detto metodo sia senz'altro preferibile per le applicazioni pratiche. Converrà esaminare, a tal fine

di caso in caso, se le ipotesi su cui esso si basa rispondono o meno a realtà. Il metodo potrà, ad esempio, plausibilmente applicarsi quando si tratta di determinare le probabilità annue di morte per le età adulte, poichè, per gli individui di tali età, i saggi di eliminazione non variano sostanzialmente entro un anno di vita, mentre può plausibilmente ritenersi che gli entrati e gli usciti siano proporzionali ai viventi.

9. — Formula di quadratura dei trapezi e di Cavalieri-Simpson per il calcolo di $\int_t^{t+1} E_s\, ds$. — Se poi il materiale statistico a disposizione, in corrispondenza ad ogni intervallo $t + \frac{i-1}{n} \vdash t + \frac{i}{n}$ (per $i = 1, 2, \ldots, n$) dà i valori di:

$$k_{t,i} = \int_{t+\frac{i-1}{n}}^{t+\frac{i}{n}} k_u\, du \qquad \text{e} \qquad m_{t,i} = \int_{t+\frac{i-1}{n}}^{t+\frac{i}{n}} m_u\, du \tag{20}$$

o, quanto meno, quelli $k_{t,i} - m_{t,i}$, per ridurre l'errore di approssimazione ottenuto usando la (19) si potrebbe supporre l'uniforme distribuzione delle differenze tra i numeri degli entrati e degli usciti o eliminati in ciascun intervallo i^{esimo} e calcolare le corrispondenti durate medie di esposizione al rischio di ciascun contingente di individui. Ad es. quello i^{esimo}, $k_{t,i} - m_{t,i}$ avrebbe durata media di esposizione di:

$$1 - \frac{2i-1}{2n} \tag{21}$$

uguale alla differenza tra l'ampiezza unitaria dell'intervallo tra t e $t+1$ e il punto medio dell'intervallino i^{esimo}.

Se, per fissare le idee, supponiamo $n=4$ — che equivarrebbe ad una rilevazione trimestrale se l'intervallo $t \vdash t+1$ fosse di un anno — dalla precedente per $i = 1, 2, 3, 4$, avremmo le

seguenti durate medie di esposizione : $\frac{7}{8}$, $\frac{5}{8}$, $\frac{3}{8}$, $\frac{1}{8}$ e quindi :

$$\int_t^{t+1} (k_u - m_u)\,(t + 1 - u)\,du \cong \frac{1}{8}\Big[7\,k_{t,1} + 5\,k_{t,2} + 3k_{t,3} + k_{t,4} - \tag{22}$$

$$- (7\,m_{t,1} + 5\,m_{t,2} + 3\,m_{t,3} + m_{t,4})\Big]$$

Questo valore di calcolo non faticoso, sostituito nella (2), permetterà di ricavare un valore approssimato di S_t, più di quanto lo sia ${}_aS_t$ dato dalla (19).

Formalmente diversa ma sostanzialmente identica è l'espressione che può ricavarsi, per $n = 4$, da quella indicata dal Cantelli per il calcolo di $\int_t^{t+1} E_s\,ds$ usando la formula di quadratura dei trapezi, supponendo l'andamento di E_s lineare tra gli estremi di due intervalli successivi (1).

Se allora ammettiamo che E_u abbia in $t \vdash t + 1$ la derivata seconda determinata e finita, l'errore nel calcolo di questo integrale (e quindi del denominatore della (2), sarà $\varepsilon \leq 1/12\,n^2 \mid E'' \mid$ dove $\mid E'' \mid$ indica un numero maggiore o uguale, in valore assoluto, a tutti i valori che la E''_s prende in $t \vdash t + 1$.

Usando invece la formula di quadratura di Cavalieri-Simpson, supponendo l'andamento di E_s parabolico tra gli estremi di due intervalli successivi, si perviene, nel caso $n = 4$, alla seguente espressione :

$$\int_t^{t+1} E_s\,ds = l_t + \frac{1}{12}\Big[11\,k_{t,1} + 7\,k_{t,2} + 5\,k_{t,3} + k_{t,4} - (11m_{t,1} + \tag{23}$$

$$+ 7\,m_{t,2} + 5\,m_{t,3} + m_{t,4})\Big]$$

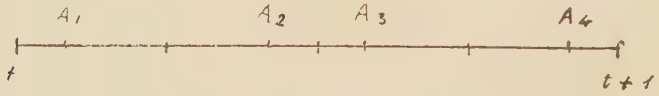
e se E_s ha derivata quarta determinata e finita, indicando con $\mid E^{IV} \mid$ il massimo valore assoluto di questa derivata nell'inter-

(1) F.P. CANTELLI, *Sulla costruzione...*, cit., pag. 24, tenendo presente che, con le nostre notazioni, è ad es. $k\left(t, t + \frac{\nu}{10}\right) = k_{t,1} + k_{t,2} + \ldots + k_{t,\nu}$.

vallo $t \vdash t + 1$, l'errore della precedente formula approssimata sarà :

$$\varepsilon_2 \leq \frac{1}{2880\, n^4} \mid E^{IV} \mid$$

I coefficienti dei $k_{t,i}$ e degli $m_{t,i}$ mostrano che l'uso della formula di quadratura di Cavalieri-Simpson equivale ad ammettere che gli individui $k_{t,i} - m_{t,i}$ $(i = 1, 2, \ldots, 4)$ abbiano una durata di esposizione rispettivamente uguale alla distanza da $t + 1$ dei punti A_1, A_2, A_3, A_4 segnati nel seguente segmento :



Questa disposizione dei punti A_i, per quanto simmetrica, è un po' arbitraria da un punto di vista concreto. Ciò fa dubitare che l'uso di formule di quadrature più complicate porti a risultati più esatti.

Nel n. 13 riporteremo delle applicazioni di queste formule.

10. — Approssimazione di $\int_t^{t+1} \mu_u \, du$ per mezzo dei saggi medi. — I precedenti procedimenti interpolatori sono serviti a determinare dei valori presumibilmente meglio approssimati di S_t.

Ora, questi saggi medi di eliminazione variano generalmente di valore al variare di t. Può allora sembrare opportuno procedere ad una conveniente interpolazione di essi per ottenere una espressione algebrica che esprima la μ_u in tutto l'intervallo che interessa, e servirsi poi di questa per ottenere l'espressione esatta di q_t in funzione dei parametri che determinano μ_u, ricordando che :

$$q_t = 1 - e^{-\int \mu_u \, du} \qquad (5)$$

la quale non comporta ipotesi restrittive.

Dobbiamo però osservare che allo scopo della determinazione di q_t con la precedente, quello che interessa non è tanto

la conoscenza della funzione μ_u nell'intervallo $t \vdash t+1$ quanto il valore di $\int_t^{t+1} \mu_u \, du$.

Perciò l'ipotesi $\mu_u = \text{cost.} = S_t$ dà lo stesso valore per q_t come una qualunque altra ipotesi più realistica, purchè l'area racchiusa dalla funzione μ_u tra t e $t+1$ sia esattamente eguale all'area del rettangolo che sullo stesso intervallo ha l'altezza uguale ad S_t.

Se la rilevazione statistica, in corrispondenza ad ogni intervallo $t + \frac{i-1}{n} \vdash t + \frac{i}{n}$ (per $i = 1, 2 \ldots\ldots n$) dà gli elementi indicati nella (20), o, quanto meno, quelli di $k_{t,i} - m_{t,i}$ ci si potrebbe mettere nelle condizioni più generali che non valga nessuna delle ipotesi precedentemente esaminate per poter calcolare q_t.

Si presenta, infatti, immediato il procedimento di *approssimare direttamente l'integrale che figura ad esponente di e* nella precedente espressione sostituendo alla curva rappresentatrice di μ_u — incognita e comunque complicata — ***una spezzata a gradinata*** supponendo che nell'intervallo generico i^{esimo}: $t + \frac{i-1}{n} \vdash t + \frac{i}{n}$ $(i = 1, 2, \ldots, n)$ il saggio istantaneo $\mu_{t,i}$ sia costante e quindi uguale a:

$$S_{t,i} = \frac{m_{t,i}}{\frac{1}{n} \, l_{t+\frac{i-1}{n}} + \int_{t+\frac{i-1}{n}}^{t+\frac{i}{n}} (k_u - m_u) \, (t + \frac{i}{n} - u) \, du} \qquad (24)$$

dove:

$$l_{t+\frac{i-1}{n}} = l_t + k_{t,1} + \ldots + k_{t,\frac{i-1}{n}} - (m_{t,1} + \ldots + m_{t,\frac{i-1}{n}})$$

ed il fattore $\frac{1}{n}$ sta a indicare la durata di esposizione al rischio di eliminazione.

Questo processo di approssimazione è, come abbiamo già detto, comunemente usato in Statistica, quando, nell'operare su seriazioni di caratteri continui, rilevati per classi di intensità, si attribuisce a tutti gli elementi di una stessa classe l'intensità media o l'intensità centrale della classe, salvo poi — nei casi in cui interessa grande precisazione — a procedere ad opportune correzioni delle costanti caratteristiche, o alla determinazione dei limiti di approssimazione, in base ad opportune ipotesi sull'andamento della funzione di frequenza (1).

Non conoscendo le funzioni k_u e m_u, procediamo ad un calcolo approssimato dell'integrale che figura nella (24). Si ha:

$$\int_{t+\frac{i-1}{n}}^{t+\frac{i}{n}} (k_u - m_u)\left(t + \frac{i}{n} - u\right) du = \left(t + \frac{i}{n}\right) \int_{t+\frac{i-1}{n}}^{t+\frac{i}{n}} (k_u - m_u)\, du - \tag{25}$$

$$\int_{t+\frac{i-1}{n}}^{t+\frac{i}{n}} u\, k_u\, du + \int_{t+\frac{i-1}{n}}^{t+\frac{i}{n}} u\, m_u\, du = \left(t + \frac{i}{n}\right)(k_{t,i} - m_{t,i}) -$$

$$k_{t,i}\, A_k(u) + m_{t,i}\, A_m(u)$$

dove:

$$A_k(u) = \frac{1}{k_{t,i}} \int_{t+\frac{i-1}{n}}^{t+\frac{i}{n}} u\, k_u\, du \qquad A_m(u) = \frac{1}{m_{t,i}} \int_{t+\frac{i-1}{n}}^{t+\frac{i}{n}} u\, m_u\, du \tag{26}$$

I valori $A_k(u)$ e $A_m(u)$ rappresentano quindi dei valori medi, e precisamente i punti interni all'intervallo $t + \frac{i-1}{n} \vdash t + \frac{i}{n}$ nel primo del quale può pensarsi che avvenga l'immissione dei

---

(1) Cfr. T. Salvemini, *Sulla correzione dei momenti empirici di una distribuzione statistica*, « Giornale dell'Istituto Italiano degli Attuari », Anno X, n. 1, 2, Roma, gennaio-aprile 1939.

$k_{t,i}$ individui nella collettività, e nel secondo che avvenga l'eliminazione degli $m_{t,i}$ individui.

Se nulla si conosce sull'eventuale posizione di questi punti medi, allora si porrà — in relazione al principio di indifferenza — $A_k(u) = A_m(u) = t + \frac{2i-1}{2n}$ cioè uguale al punto medio di detto intervallo, e l'integrale precedente diventa uguale a $\frac{k_{t,i} - m_{t,i}}{2n}$ [1].

Se, invece, si ritiene che ad es. i $k_{t,i}$ individui si addensino prevalentemente nella seconda metà di detto intervallo, mentre le eliminazioni per la causa c avvengano prevalentemente nella prima metà, allora converrà porre $A_k(u) = t + \frac{i-1}{n} + \frac{3}{4}\frac{1}{n}$, $A_m(u) = t + \frac{i-1}{n} + \frac{1}{4}\frac{1}{n}$ e il precedente integrale diverrà uguale a $\frac{1}{n}\left(\frac{1}{4}k_{t,i} - \frac{3}{4}m_{t,i}\right)$.

Nei casi concreti, in base alla conoscenza che si può avere di questo movimento di entrata, di uscita e di eliminazione, si potrebbero prendere tanto per $A_k(u)$ che per $A_m(u)$ due diverse posizioni, delle quali, una si giudichi approssimata in un senso e l'altra approssimata nell'altro senso. In tal modo si otterranno per q_t due valori approssimati al valore esatto in senso opposto, e potrà vedersi, facilmente, l'influenza delle supposizioni fatte, ed avere, così, degli elementi per giudicare sui valori esatti di q_t.

Limitandoci a scrivere l'espressione del saggio medio nel caso che nulla si conosca sulla posizione dei punti medi suddetti, si avrà:

$$ {}_aS_{t,i} = \frac{m_{t,i}}{\frac{1}{n}l_{t+\frac{i-1}{n}} + \frac{k_{t,i} - m_{t,i}}{2n}} \qquad (27) $$

---

[1] Ciò porta evidente a un valore approssimato. Affinchè sia esatto, compatibilmente con l'ipotesi μ = cost. nei singoli intervalli, dovrebbe valere un certo andamento per $k_{t,i}$ (Cfr. CANTELLI, *Sulla costruzione...*, cit.). Se questo non vale, e generalmente non vale, vuol dire che siamo nel campo delle approssimazioni, rese necessarie dall'impossibilità di avere il valore esatto dell'integrale in considerazione.

e quindi, assumendo questa quantità come saggio costante nello intervallo $t + \frac{i-1}{n} \vdash t + \frac{i}{n}$, si avrà, approssimativamente :

$$\int_{t+\frac{i-1}{n}}^{t+\frac{i}{n}} \mu_s \, d\, s \simeq \frac{1}{n} {}_aS_{t,i} = \frac{m_{t,i}}{l_{t+\frac{i-1}{n}} + \frac{k_{t,i} - m_{t,i}}{2}} \qquad (28)$$

La differenza tra il primo e l'ultimo membro della precedente dipende dalle due approssimazioni dette : cioè dall'aver sostituito in detto intervallo al saggio variabile μ_s il saggio medio $S_{t,i}$ e per aver preso come valore di tale saggio la quantità ${}_aS_{t,i}$. Può darsi che queste due approssimazioni si sommino ma può anche darsi che si sottraggano o che proprio si compensino.

Facendo la somma delle precedenti quantità per tutti gli intervalli considerati si avrà :

$$\int_t^{t+1} \mu_s \, ds \simeq \frac{1}{n} \sum_{i=1}^{n} {}_aS_{t,i} \simeq \sum_{i=1}^{n} \frac{m_{t,i}}{l_{t+\frac{i-1}{n}} + \frac{k_{t,i} - m_{t,i}}{2}} \qquad (28')$$

e quindi, come valore approssimato della probabilità di eliminazione nell'intervallo $t \vdash t + 1$, potremo assumere la seguente espressione, di facile calcolo :

$$q'_t = 1 - e^{-\frac{1}{n} \sum_{i=1}^{n} {}_aS_{t,i}} \qquad (29)$$

Poichè è da ritenere che le differenze tra il primo e l'ultimo membro della (28) non siano tutte dello stesso segno perchè se una tale sistematicità si avesse in realtà per ogni intervallo parziale considerato, essa non dovrebbe sfuggire allo statistico, il secondo membro della (28') darà presumibilmente, nei casi concreti, e nella maggior parte degli intervalli, un valore prossimo al valore esatto dell'integrale che figura nel primo membro. È chiaro però che una migliore approssimazione si ottiene generalmente prendendo i valori esatti $S_{t,i}$ e non quelli approssimati ${}_aS_{t,i}$. In ogni caso abbiamo visto che un errore uguale ad α nella va-

lutazione di $\int_t^{t+1} \mu_s \, ds$ comporta per q'_t un errore inferiore ad α, e tanto più piccolo di α quanto più è alto il valore di detto integrale.

Si rilevi, inoltre, che mentre per la determinazione del saggio medio di eliminazione nell'intervallo generico $t \vdash t+1$ è opportuno eseguire una media ponderata dei saggi relativi agli intervalli parziali — come già rilevammo al n. 4, ragionando nel continuo — per la determinazione di un valore approssimato della quantità $\int_t^{t+1} \mu_s \, ds$ che figura nella (7) — e che non mi pare che abbia un significato demografico a sè — è opportuno eseguire una media aritmetica dei saggi corrispondenti agli intervalli parziali. A meno che, essendo molto esiguo il numero dei casi osservati in qualcuno degli intervalli parziali, si ritenga che il corrispondente $S_{t,i}$ sia molto influenzato da questo numero limitato di casi, per cui si ravvisi più opportuno limitarne l'influenza ricorrendo a una media ponderata.

Infine, tenendo presente che gli $S_{t,i}$ possono pensarsi come valori approssimati di μ_s in corrispondenza ai punti centrali degli intervalli $t + \frac{i-1}{n} \vdash t + \frac{i}{n}$, si potrebbero usare anche altre formule di quadratura ai fini del calcolo di $\int_t^{t+1} \mu_s \, ds$.

Così ad es. si potrebbe pensare di far passare una parabola del 5° ordine $P_5(t)$ per i sei punti aventi per ascisse rispettivamente i valori $t - \frac{1}{8}$, $t + \frac{1}{8}$, $t + \frac{3}{8}$, $t + \frac{5}{8}$, $t + \frac{7}{8}$, $t + 1 + \frac{1}{8}$ e per ordinate rispettivamente i valori $S_{t-1,4}$, $S_{t,1}$, $S_{t,2}$, $S_{t,3}$, $S_{t,4}$, $S_{t+1,1}$ e poi prendere come valore approssimato di $\int_t^{t+1} \mu_s \, ds$ quello di $\int_t^{t+1} P_5(x) \, dx$.

Essendo però generalmente sconosciuto il tipo di funzione che si adatta esattamente ai μ_s, non sarà possibile determinare l'errore che si commette anche con tali altre formule approssimate.

Pertanto, nei casi concreti in cui non sia possibile precisare l'andamento della legge di eliminazione incognita, in base al principio di indifferenza e a quello della massima semplicità, riteniamo opportuno seguire il metodo sopra esposto di fare la media aritmetica degli $S_{t,i}$, ciò che equivale ad usar la formula di quadratura dei trapezi circoscritti per il calcolo di $\int_t^{t+1} \mu_s \, ds$.

Osserviamo, pure, che la conoscenza degli S_t (o quella degli $S_{t,i}$) dà la possibilità di determinare *l'area di probabile giacitura* o striscia di indeterminazione della distribuzione ignota dei μ_s. È chiaro, infatti, che considerando gli S_t come valori approssimati di μ_s, una prima volta in corrispondenza all'ascissa t e una seconda volta in corrispondenza all'ascissa $t + 1$, si potranno determinare, con procedimento grafico o con procedimento analitico, due curve che limitano la detta area o striscia. È opportuno però che la successione degli S_t non presenti delle brusche oscillazioni di carattere accidentale; nel caso, si può procedere prima ad una conveniente perequazione (ad es. grafica o per medie). Si possono calcolare così valori approssimati per difetto e valori approssimati per eccesso di $\int \mu_s \, ds$.

Dalla conoscenza di queste curve si potrà poi passare a quelle che limitano la distribuzione ignota dei q_t.

Ciò è utile perchè richiama l'attenzione dello studioso, o dell'attuario che deve adoperare i q_t per fini concreti, sull'inevitabile approssimazione di queste probabilità e quindi a usare prudenza nell'interpretazione dei risultati.

Se ragioni particolari possono far ritenere consigliabile di prendere, per ogni t, la media aritmetica dei corrispondenti valori approssimati per difetto e per eccesso, non si avrebbe più una ipotesi diretta sulla legge incognita di eliminazione, ma semplicemente un metodo pratico per determinare, approssimativamente, i valori di q_t. L'esame del fenomeno e le sue finalità potranno essere di guida nella scelta del procedimento migliore da seguire.

11. — Grado di approssimazione dei procedimenti indicati in un esempio concreto. — Allo scopo di esaminare, in concreto, il grado di approssimazione che si raggiunge con i procedimenti esposti nei n. 9 e 10, ci siamo valsi dello stesso mate-

riale elaborato dal prof. Gini — e messoci gentilmente a disposizione — relativo alle coniugate delle Case sovrane, d'Europa, delle Case mediatizzate tedesche, delle Case principesche d'Italia, Francia, Germania, Austria, Ungheria, Russia, Polonia e Gran Bretagna, avendosi per ciascuna coniugata la data precisa delle nozze, quella di nascita dei figli e quella di scioglimento del matrimonio o per morte propria o per morte del marito o per divorzio o per fine del periodo di osservazione.

I contingenti di eliminati, di entrati e di usciti sono stati classificati secondo i successivi trimestri di età.

Riunendo per anno i risultati del nostro spoglio trimestrale, abbiamo avuto, in qualche caso, lievissime divergenze con gli l_t, m_t e k_t pubblicati dal Gini nella tav. XIV del suo lavoro cit.; riteniamo che ciò debba attribuirsi al fatto che nel fascicolo avuto non siamo riusciti a rintracciare alcune tabelle in cui erano riassunti i dati delle schede originarie del Savorgnan, per cui abbiamo dovuto desumere quelli relativi ad alcune Case principesche da altre tavole di spoglio (1).

Poichè questo esempio ha, come abbiamo detto, uno scopo esemplificativo dei procedimenti indicati, piuttosto che procedere alla ricerca dei dati originarii che mancavano per la nostra classificazione, abbiamo creduto più rapido, per i nostri confronti, ricalcolare i valori di S_t tenendo conto delle durate esatte di esposizione al rischio, quali risultano dal nostro spoglio. È da notare subito che tra gli S_t ricalcolati e quelli pubblicati nel lavoro del Gini ci sono differenze trascurabili.

Eseguiamo i calcoli da 19 anni in poi, perchè lo scarso numero di casi prima di questa età porta a valori fortemente perturbati e quindi poco significativi.

La Tav. II che segue riporta i vari elementi per il calcolo di ${}_aS_{t,i}$ in base alla (27). Nell'ultima colonna si sono riportati i valori esatti di $S_{t,i}$ in base al computo — possibile con i dati specifici in nostro possesso — delle durate effettive di esposi-

---

(1) Le tabelle mancanti sono state ritrovate successivamente in un altro fascicolo.

I lavori del prof. F. Savorgnan sulla nuzialità e la fecondità delle aristocrazie sono pubblicati in «Metron», Volumi III, IV, V e IX.

TAV. II.

Elementi per il calcolo dei saggi trimestrali di eliminazione ${}_aS_{t,i}$.

Età	Spose che alle nozze hanno età tra t e $t+¼$		Spose che al concepimento hanno età tra t e $t+¼$		Spose che 12 mesi dopo un concepimento hanno età tra t e $t+¼$		Spose uscite di osservazione o per scioglimento del matr. o per fine del periodo di osservazione		$\sum_{i=0}^{t,i-1}(s_{t,i}-m_{t,i}+r_{t,i}-f_{t,i}-e_{t,i})$		$s_{t,i}+r_{t,i}-(f_{t,i}+e_{t,i})$		$l_{t,i}+\frac{k_{t,i}-m_{t,i}}{2}$	$\frac{m_{t,i}}{{}_4aM_{t,i}}$	$\frac{m_{t,i}}{4M_{t,i}}$
t,i	$s_{t,i}$		$m_{t,i}$		$r_{t,i}$		$f_{t,i}+e_{t,i}$		$l_{t,i}$		$k_{t,i}$		${}_4aM_{t,i}$	¼ ${}_aS_{t,i}$	¼ $S_{t,i}$
< 19	54		31		8		—		—		62		—	—	—
19	18		9		4		—		31		22		37,5	0,2400	0,2515
$19^1/_4$	21	66	17	53	8	23	1	2	44	31	28	87	49,5	3434	0,3340
$19^2/_4$	12		12		6		—		55		18		58	2069	2080
$19^3/_4$	15		15		5		1		61		19		63	2381	2360
20	17		19		9		2		65		24		67,5	2815	2836
$20^1/_4$	18		16		17		—		70		35		79,5	2013	2057
$20^2/_4$	23	77	21	78	12	53	1	4	89	65	34	126	95,5	2199	2079
$20^3/_4$	19		22		15		1		102		33		107,5	2046	2061
21	24		26		19		1		113		42		121	2149	2184
$21^1/_4$	20		23		16		—		129		36		135,5	1697	1679
$21^2/_4$	15	77	35	116	21	78	—	1	142	113	36	154	142,5	2456	2444
$21^3/_4$	18		32		22		—		143		40		147	2177	2224
22	12		28		26		—		151		38		156	1795	1743
$22^1/_4$	14	60	21	108	23	116	1	2	161	151	36	174	168,5	1240	1226
$22^2/_4$	24		31		35		1		176		58		189,5	1636	1624
$22^3/_4$	10		28		32		—		203		42		210	1333	1316
23	12		32		28		2		217		38		220	1455	1473
$23^1/_4$	12		30		21		2		223		31		223,5	1342	1362
$23^2/_4$	7	39	32	120	31	108	1	6	224	217	37	141	226,5	1413	1441
$23^3/_4$	8		26		28		1		229		35		233,5	1113	1090
24	21		41		32		—		238		53		244	1680	1670
$24^1/_4$	10		31		30		—		250		40		254,5	1218	1213
$24^2/_4$	14	54	35	139	32	120	1	3	259	238	45	171	264	1326	1323
$24^3/_4$	9		32		26		2		269		33		263,5	1187	1180

Elementi per il calcolo dei saggi trimestrali di eliminazione ${}_aS_{t,i}$.

Età	Spose che alle nozze hanno età tra t e $t+¼$	Spose che al concepimento hanno età tra t e $t+¼$	Spose che 12 mesi dopo un concepimento hanno età tra t e $t+¼$	Spose uscite di osservazione o per scioglimento del matr. o per fine del periodo di osservazione	$\sum_{z=0}^{t,i-1}(s_{t,z}-m_{t,z}+r_{t,z}-f_{t,z}-e_{t,i})$	$s_{t,i}+r_{t,i}-(f_{t,i}+e_{t,i})$	$l_{t,i}+\frac{k_{t,i}-m_{t,i}}{2}$	$\frac{m_{t,i}}{4_aM_{t,i}}$	$\frac{m_{t,i}}{4.M_{t,i}}$
t,i	$s_{t,i}$	$m_{t,i}$	$r_{t,i}$	$f_{t,i}+e_{t,i}$	$l_{t,i}$	$k_{t,i}$	${}_4M_{t,i}$	¼ ${}_as_{t,i}$	¼ $s_{t,i}$
25	11	25	41	3	270	49	282	0,0887	0,0899
$25^1/_4$	18	36	31	2	294		299,5	1202	1213
$25^2/_4$	5	28	35	1	305	59	310,5	0902	0915
$25^3/_4$	7	40	32	1	316	38	315	1270	1290
26	4	24	25	3	314	26	315	0772	0757
$26^1/_4$	3	32	36	1	316	38	319	1003	1001
$26^2/_4$	10	40	28	—	322	38	321	1246	1230
$26^3/_4$	9	34	40	1	320	48	327	1040	1031
27	9	32	24	1	334	32	334	0958	0964
$27^1/_4$	4	27	32	1	334	35	338	0799	0797
$27^2/_4$	10	36	40	1	342	49	348,5	1033	1036
$27^3/_4$	3	22	34	2	355	35	361,5	0609	0612
28	7	28	32	2	368	37	372,5	0752	0746
$28^1/_4$	1	22	27	2	377	26	379	0580	0583
$28^2/_4$	5	31	36	2	381	39	385	0805	0808
$28^3/_4$	9	28	22	1	389	30	390	0718	0715
29	5	40	28	2	391	31	386,5	1035	1035
$29^1/_4$	5	24	22	2	382	25	382,5	0627	0626
$29^2/_4$	6	23	31	—	383	37	390	0590	0584
$29^3/_4$	4	25	28	1	397	31	400	0625	0624
30	3	24	40	1	403	42	412	0583	0585
$30^1/_4$	1	21	24	2	421	23	422	0498	0499
$30^2/_4$	3	22	23	1	423	25	424,5	0518	0520
$30^3/_4$	9	28	25	1	426	33	428,5	0653	0653

	41		139	7	70	173			
	26	129		5	314	150			
	26	130	129						
	26	117	130	5	334	151			
	22	109	117	7	368	132			
	20	112	109	5	391	124			
	16	95	112	5	403	123			

Elementi per il calcolo dei saggi trimestrali di eliminazione ${}_aS_{t,i}$.

Età	Spose che alle nozze hanno età tra t e $t+¼$	Spose che al concepimento hanno età tra t e $t+¼$	Spose che 12 mesi dopo un concepimento hanno età tra t e $t+¼$	Spose uscite di osservazione o per scioglimento del matr. o per fine del periodo di osservazione	$\sum_{i=0}^{t,i-1}(s_{t,i}-m_{t,i}+r_{t,i}-f_{t,i}-e_{t,i})$	$s_{t,i}+r_{t,i}-(f_{t,i}+e_{t,i})$	$l_{t,i}+\frac{k_{t,i}-m_{t,i}}{2}$	$\frac{m_{t,i}}{4_aM_{t,i}}$	$\frac{m_{t,i}}{4.M_{t,i}}$
t,i	$s_{t,i}$	$m_{t,i}$	$r_{t,i}$	$f_{t,i}+e_{t,i}$	$l_{t,i}$	$k_{t,i}$	${}_4aM_{t,i}$	$¼\ aS_{t,i}$	$¼\ S_{t,i}$
31	1	20	24	2	431	23	432,5	0,0462	0,0462
$31^1/_4$	3 \} 12	24 \} 92	21 \} 95	3 \} 13	434 \} 431	21 \} 94	435,5	0551	0554
$31^2/_4$	7	25	22	3	431	26	431,5	0579	0582
$31^3/_4$	1	23	28	5	432	24	432,5	0532	0529
32	2	27	20	1	433	21	430	0628	0626
$32^1/_4$	3 \} 11	15 \} 75	24 \} 92	3 \} 10	427 \} 433	24 \} 93	431,5	0348	0346
$32^2/_4$	1	11	25	1	436	25	443	0248	0249
$32^3/_4$	5	22	23	5	450	23	450,5	0488	0491
33	1	12	27	1	451	27	458,5	0262	0262
$33^1/_4$	4 \} 8	8 \} 53	15 \} 75	3 \} 12	466 \} 451	16 \} 71	470	0170	0170
$33^2/_4$	1	14	11	4	474	8	471	0297	0296
$33^3/_4$	2	19	22	4	468	20	468,5	0406	0405
34	1	18	12	1	469	12	466	0386	0387
$34^1/_4$	— \} 7	19 \} 66	8 \} 53	2 \} 6	463 \} 469	6 \} 54	456,5	0416	0419
$34^2/_4$	2	15	14	1	450	15	450	0333	0335
$34^3/_4$	4	14	19	2	450	21	453,5	0300	0300
35	2	10	18	3	457	17	460,5	0217	0216
$35^1/_4$	— \} 3	10 \} 41	19 \} 66	7 \} 20	464 \} 457	12 \} 49	465	0215	0215
$35^2/_4$	—	12	15	1	466	14	467	0257	0256
$35^3/_4$	1	9	14	9	468	6	466,5	0193	0192
36	1	8	10	4	465	7	464,5	0172	0173
$36^1/_4$	— \} 2	18 \} 53	10 \} 41	4 \} 17	464 \} 465	6 \} 26	458	0393	0392
$36^2/_4$	—	12	12	4	452	8	450	0267	0267
$36^3/_4$	1	15	9	5	448	5	443	0338	0338

Età	Spose che alle nozze hanno età tra t e $t+¼$	Spose che al concepimento hanno età tra t e $t+¼$	Spose che 12 mesi dopo un concepimento hanno età tra t e $t+¼$	Spese uscite di osservazione o per scioglimento del matr. o per fine del periodo di osservazione	$\sum_{i=0}^{t,i-1}(s_{t,i}-m_{t,i}+r_{t,i}-f_{t,i}-e_{t,i})$	$s_{t,i}+r_{t,i}-(f_{t,i}+e_{t,i})$	$l_{t,i}+\frac{k_{t,i}-m_{t,i}}{2}$	$\frac{m_{t,i}}{4_a M_{t,i}}$	$\frac{m_{t,i}}{4.M_{t,i}}$
t,i	$s_{t,i}$	$m_{t,i}$	$r_{t,i}$	$f_{t,i}+e_{t,i}$	$l_{t,i}$	$k_{t,i}$	$4_a M_{t,i}$	$¼\ a s_{t,i}$	$¼\ s_{t,i}$
37	—	6	8	4	438	4	437	0,0137	0,0137
$37^1/_4$	—	6	18	3	436	15	440,5	0136	0137
$37^2/_4$	—	11	12	3	445	9	444	0248	0249
$37^3/_4$	1	13	15	1	443	15	444	0293	0294
38	2	14	6	6	445	2	439	0319	0318
$38^1/_4$	1	3	6	8	433	−1	431	0070	0070
$38^2/_4$	1	9	11	4	429	8	428,5	0210	0210
$38^3/_4$	2	6	13	2	428	13	431,5	0139	0138
39	1	3	14	5	435	10	438,5	0068	0068
$39^1/_4$	2	9	3	9	442	4	435,5	0207	0206
$39^2/_4$	—	4	9	4	429	5	429,5	0093	0094
$39^3/_4$	—	8	6	3	430	3	427,5	0187	0187
40	—	4	3	7	425	−4	421	0095	0095
$40^1/_4$	1	2	9	4	417	6	419	0048	0048
$40^2/_4$	—	2	4	7	421	−3	418,5	0048	0048
$40^3/_4$	4	2	8	4	416	8	419	0048	0048
41	3	4	4	7	422	0	420	0095	0096
$41^1/_4$	—	5	2	5	418	−3	414	0121	0121
$41^2/_4$	—	2	2	7	410	−5	406,5	0049	0049
$41^3/_4$	—	5	2	9	403	−7	397	0126	0126
42	2	—	4	4	391	2	392	—	—
$42^1/_4$	1	3	5	7	393	−1	391	0077	0077
$42^2/_4$	—	2	2	1	389	1	388,5	0051	0051
$42^3/_4$	—	—	5	3	388	2	389	—	—

zione al rischio di concepimento del gruppo di spose in osservazione. Economie di spazio, non permettono, evidentemente, di riprodurre le successive tappe che portano agli $M_{t,i}$; il procedimento è quello del Gini.

Le differenze tra $S_{t,i}$ ed ${}_aS_{t,i}$ sono alcune volte sensibili — nella seconda cifra decimale — ma sommando i quattro valori corrispondenti ai trimestri di ciascun anno di età, ai fini della determinazione dei saggi e probabilità annuali di eliminazione per concepimento, dette differenze in parte si compensano, tranne per qualche anno in cui gli ${}_aS_{t,i}$ relativi ai trimestri sono tutti inferiori o tutti superiori ai corrispondenti $S_{t,i}$.

Dall'ultima colonna della tavola II si vede inoltre che le sensibili oscillazioni degli $S_{t,i}$ sconsigliano di ricorrere al metodo di interpolazione prospettato alla fine del n. 10 per il calcolo approssimato di $\int_t^{t+1} \mu_s \, ds$ giacchè esso richiederebbe per la funzione μ_s un andamento molto irregolare, certamente lontano dal vero.

L'osservazione degli $m_{t,i}$ e dei $k_{t,i}$ nei successivi trimestri non mostra alcuna regolarità utile allo scopo di determinare la legge di variazione di tali contingenti nel corso del trimestre o nel corso dell'anno.

Supposto di concentrare a metà trimestre questi contingenti, possiamo determinare i punti medi dell'intervallo annuale nei quali possiamo pensare che avvenga l'immissione dei k_t (annuali) e l'eliminazione degli m_t (pure annuali). Riportiamo questi risultati per vedere di quanto ci si scosta, nella formula (19), dal valore $\frac{1}{2}$ ivi considerato. Prendendo le distanze dal punto $t + 1$, useremo la formula:

$$A_k = \frac{7\,k_{t,1} + 5\,k_{t,2} + 3\,k_{t,3} + k_{t,4}}{8\,(k_{t,1} + k_{t,2} + k_{t,3} + k_{t,4})}$$

e una formula analoga (sostituendo i $k_{t,i}$ con gli $m_{t,i}$) per avere l'altro punto medio A_m.

I risultati ottenuti (approssimati rispetto ai loro valori esatti che si ottengono con le (26) del n. 10) mostrano che le differenze dal valore 0,5 preso nella (19) non sono affatto notevoli, tranne per gli A_k in alcune età alte; in questi casi, però, l'entità dei

$k_{t,i}$ è piccola in confronto agli $l_{t,i}$ per cui non si ha, in definitiva, un errore sensibile.

Età	Ak	Am
19	0,53	0,47
20	0,46	0,47
21	0,50	0,47
22	0,48	0,50
23	0,50	0,52
24	0,54	0,52
25	0,53	0,46
26	0,45	0,46
27	0,48	0,52
28	0,51	0,49
29	0,49	0,55
30	0,53	0,48
31	0,49	0,49
32	0,49	0,53
33	0,55	0,44
34	0,42	0,53
35	0,58	0,50
36	0,57	0,46
37	0,42	0,41
38	0,26	0,57
39	0,61	0,45
40	0,16	0,57
41	0,33	0,46
42	0,69	0,53

Nella Tav. III riportiamo i valori di $\frac{1}{4} \Sigma S_{t,i}$ e $\frac{1}{4} \Sigma {}_aS_{t,i}$ dove la somma è estesa ai quattro valori relativi ai trimestri di ciascun anno. In questa stessa tavola riportiamo inoltre i valori ricalcolati di S_t, gli ${}_aS_t$, ottenuti applicando la formula approssimata (19), e poi gli $S_t^{(*)}$ ed $S_t^{(c)}$ ottenuti applicando rispettivamente le formule (22) e (23) del n. 9.

È evidente che, tra essi, il valore che ad ogni età t dobbiamo ritenere più esatto come saggio medio è S_t riportato nella 2ª colonna. –

I valori di $S_t^{(*)}$ ed $S_t^{(c)}$ sono alquanto prossimi ai valori di S_t. La differenza è soprattutto notevole in corrispondenza a $t = 22$ anni, con valori approssimati per eccesso, e a $t = 25$ anni, con valori approssimati per difetto. Guardando i dati più analitici contenuti nella Tav. II rileviamo che, in particolare, in cor-

rispondenza a 22 anni i $k_{t,i}$ del 3 e 4 trimestre sono sensibilmente superiori a quelli del 1 e 2 trimestre, mentre il contrario avviene per i $k_{t,i}$ dei trimestri del 25° anno di età; a questa scarsa uniformità del movimento di uscita e di ingresso è certo da attribuire le differenze riscontrate.

TAV. III.

Saggi medi di eliminazione ottenuti con vari procedimenti

Età	s_t	$\frac{1}{4}\Sigma\, s_{t,i}$	$\frac{1}{4}\Sigma_a s_{t,i}$	$s_t^{(*)}$	$s_t^{(c)}$	${}_a s_t$
1	2	3	4	5	6	7
19	1,0200	1,0295	1,0284	1,0192	1,0159	1,1042
20	0,8849	0,9033	0,9073	0,8914	0,8966	0,8764
21	0,8541	0,8531	0,8479	0,8498	0,8509	0,8788
22	0,5870	0,5909	0,6010	0,5967	0,5956	0,5870
23	0,5345	0,5366	0,5323	0,5312	0,5312	0,5275
24	0,5368	0,5386	0,5411	0,5388	0,5377	0,5472
25	0,4328	0,4317	0,4261	0,4275	0,4260	0,4418
26	0,4026	0,4019	0,4061	0,4056	0,4067	0,4012
27	0,3396	0,3409	0,3399	0,3386	0,3390	0,3333
28	0,2853	0,2852	0,2855	0,2856	0,2853	0,2872
29	0,2865	0,2869	0,2877	0,2874	0,2874	0,2821
30	0,2257	0,2257	0,2252	0,2253	0,2249	0,2278
31	0,2126	0,2127	0,2124	0,2128	0ü2127	0,2130
32	0,1709	0,1712	0,1712	0,1709	0,1710	0,1697
33	0,1134	0,1133	0,1135	0,1135	0,1135	0,1152
34	0,1450	0,1450	0,1444	0,1446	0,1446	0,1425
35	0,0880	0,0879	0,0882	0,0882	0,0880	0,0905
36	0,1168	0,1170	0,1170	0,1167	0,1165	0,1173
37	0,0817	0,0817	0,0814	0,0814	0,0815	0,0827
38	0,0739	0,0736	0,0738	0,0738	0,0739	0,0726
39	0,0555	0,0555	0,0555	0,0555	0,0553	0,0558
40	0,0238	0,0239	0,0239	0,0238	0,0240	0,0236
41	0,0392	0,0392	0,0391	0,0389	0,0389	0,0392
42	0,0128	0,0128	0,0128	0,0128	0,0128	0,0128

Per passare dai saggi medi annuali alle corrispondenti probabilità useremo al solito la (5), dove come esponente di e useremo rispettivamente i valori delle varie colonne della Tav. III.

Poichè la sostituzione dell'integrale con i detti valori implica una ipotesi di cui non si può misurare con esattezza la portata dell'approssimazione che si ha nei casi concreti, non è lecito stabilire a priori quale delle probabilità indicate nella Tav. IV sia la più esatta.

Se teniamo presente che gli $S_{t,i}$ di cui all'ultima colonna della Tav. II sono valori esatti e che l'approssimazione che si ottiene per $\int_t^{t+1} \mu_s \, ds$ è da ritenersi in generale migliore prendendo $1/4 \sum_{i=1}^{4} S_{t,i}$ anzichè S_t propendiamo nel considerare più rispondenti al nostro scopo i valori della colonna 3 anzichè quelli della colonna 2.

TAV. IV.

Valori delle probabilità q_t calcolate con la formola $1 - e^{-\int_t^{t+1} \mu_s ds}$ dove al posto dell'integrale si sono sostituiti i valori approssimati della Tav. III.

Età	q_t	q'_t	$_aq'_t$	q_t^*	$q_t^{(c)}$	$_aq_t$
1	2	3	4	5	6	7
19	0,6394	0,6428	0,6424	0,6391	0,6379	0,6685
20	5878	5948	5964	5899	5920	5837
21	5743	5680	5717	5725	5730	5847
22	4440	4462	4517	4494	4488	4440
23	4140	4153	4127	4121	4121	4099
24	4154	4164	4179	4166	4159	4214
25	3513	3506	3469	3479	3469	3571
26	3324	3309	3338	3334	3341	3305
27	2880	2889	2882	2872	2875	2835
28	2482	2481	2484	2484	2482	2496
29	2491	2494	2500	2498	2498	2458
30	2020	2020	2016	2017	2014	2037
31	1915	1916	1913	1917	1916	1918
32	1571	1573	1573	1571	1572	1561
33	1072	1071	1073	1072	1072	1088
34	1350	1350	1345	1346	1346	1328
35	0842	0841	0844	0844	0842	0865
36	1102	1104	1104	1101	1100	1105
37	0784	0784	0782	0782	0783	0794
38	0712	0711	0711	0711	0712	0700
39	0540	0540	0540	0540	0538	0543
40	0235	0236	0235	0235	0237	0233
41	0384	0384	0382	0381	0381	0384
42	0127	0127	0127	0127	0127	0127

Per un più preciso confronto tra i valori delle probabilità ottenute con i vari procedimenti seguiti, calcoliamo le differenze

relative in base alla formula

$$d_r = \frac{a - A}{2A(1 - A)}$$

dove A è il valore preso per riferimento o come fondamentale ed a il valore che con esso si mette a confronto [1].

Queste differenze sono calcolate prendendo per riferimento una prima volta le probabilità q_t della col. 2 della Tav. IV e una seconda volta le probabilità q_t' della col. 3. Esse sono riportate, moltiplicate per 10.000, nella Tav. V. Di esse si è fatto, per comodità, di confronto anche la somma dei valori assoluti.

Risulta che tali differenze sono più elevate nelle prime età e diventano poco intense nelle età successive, dove si hanno sovente differenze nulle. L'alternarsi dei segni nelle colonne mostra chiaramente che non ci sono differenze sistematiche tra i vari metodi seguiti e quelli di riferimento.

Tra i q_t e i q_t' le differenze relative sono in generale basse, tranne per il secondo e terzo valore; la somma di dette differenze per questi due valori comprende quasi la metà della somma totale dei valori assoluti dalle differenze relative.

Il metodo che porta a valori delle probabilità più prossimi ai q_t è quello ricavato dall'applicazione della (22) che presuppone l'uniforme distribuzione degli entrati e degli usciti in ciascun intervallo parziale, o l'applicazione della formula di quadratura dei trapezi per il calcolo di $\int_t^{t+1} \mu_s \, ds$; per questo metodo la somma dei valori assoluti delle differenze relative è 516.

Il metodo, invece, che porta a valori delle probabilità più prossimi ai q_t' è quello che si ottiene ricorrendo alla (28') insieme alla (29), metodo che presuppone, come il precedente, l'uniforme distribuzione degli entrati e degli usciti in ciascun intervallo parziale; la somma dei valori assoluti delle differenze relative è 637.

---

[1] C. Gini, *Sul massimo degli indici di variabilità assoluta e sulle sue applicazioni agli indici di variabilità relativa e al rapporto di concentrazione*, «Metron», Vol. VIII, 1929.

Perciò, se si ritengono i valori q_t' più esatti dei q_t, e sempre che non si possano determinare le durate esatte di esposizione al rischio degli individui che fanno parte della collettività che si prende in considerazione, possiamo dire che, subordinatamente alle stesse ipotesi, il metodo suggerito al n. 10, il quale si serve

TAV. V.

Differenze relative $\frac{a - A}{2A(1 - A)}$, *moltiplicate per* 10.000, *tra i valori delle probabilità di eliminazione di cui alle Tav. IV.*

Età	Differenze relative, moltiplicate per 10.000, tra										
	q'_t e q_t	aq_t' e q_t	q^*_t e q_t	$q^{(c)}_t$ e q_t	aq_t e q_t	q_t e q_t'	aq_t' e q'_t	q^*_t e q'_t	$q^{(c)}_t$ e q'_t	aq_t e q'_t
19	74	65	— 7	— 33	631	— 74	— 9	— 81	— 107	560
20	144	177	43	87	— 85	— 145	— 33	— 102	— 58	— 230
21	— 129	— 53	— 37	— 27	213	128	75	92	102	341
22	44	156	109	97	—	— 45	— 111	65	53	— 45
23	27	— 27	— 39	— 39	— 85	— 27	— 54	— 66	— 66	— 111
24	21	51	25	10	124	— 21	31	4	— 10	— 103
25	— 15	— 97	— 75	— 97	127	15	— 81	— 59	— 81	143
26	— 34	32	23	38	— 43	34	65	56	72	— 9
27	21	5	— 20	— 12	— 110	— 22	— 17	— 41	— 34	— 131
28	— 3	5	5	—	38	3	8	8	3	40
29	8	24	19	19	— 88	— 8	16	11	11	— 96
30	—	— 12	— 9	— 19	53	—	— 12	— 9	— 19	53
31	3	— 6	6	3	10	— 3	— 10	3	—	6
32	8	8	—	4	— 38	— 8	—	— 8	— 4	— 45
33	— 5	4	—	—	72	5	10	5	5	89
34	—	— 22	— 18	— 18	— 98	—	— 21	— 17	— 17	— 94
35	— 6	13	13	—	149	7	— 21	21	7	167
36	10	10	— 5	— 10	15	— 10	—	— 15	— 20	5
37	—	— 14	— 14	— 7	69	—	— 14	— 14	— 7	69
38	— 8	— 8	— 8	—	— 91	7	—	—	7	— 83
39	—	—	—	— 20	29	—	—	—	— 20	29
40	22	—	—	44	— 44	— 22	— 22	— 22	22	— 65
41	—	— 27	— 41	— 41	—	—	— 27	— 41	— 41	—
42	—	—	—	—	—	—	—	—	—	—
$\Sigma\|d_t\|$	583	816	516	625	2.212	584	637	740	766	2.514

di un valore approssimato di $\int_t^{t+1} \mu_s \, ds$, rende meglio di quello che si serve d'un valore approssimato di $\int_t^{t+1} E_s \, ds$ per il calcolo di S_t da cui risalire poi alle probabilità q_t mediante la (6).

Anche il metodo che ricorre alla formula di quadratura di Cavalieri, per approssimare l'ultimo integrale suddetto, è meno buono del precedente, riconfermando quanto già osservammo alla fine del n. 9.

Risultati poco soddisfacenti, rispetto ai precedenti, dà quello che usa la (19), ma bisogna tener presente che esso si vale soltanto della classificazione annua e non di quella trimestrale del movimento di entrata e di uscita dalla collettività.

Conformemente a quanto abbiamo detto alla fine del n. 10, possiamo determinare l'area di probabile giacitura di μ_s (per ogni valore di s). A tale scopo ci si potrebbe servire del più piccolo e del più grande valore ottenuto per S_t, per ogni valore di t, con i vari procedimenti sopra indicati (Cfr. Tav. III, col. 3-7), ma potendosi dare il caso che, per qualche valore di t, gli S_t siano tutti approssimati per difetto o tutti approssimati per eccesso è meglio calcolare un valore per difetto di S_t concentrando i $k_{t,i}$, di cui alla Tav. II, all'inizio di ciascun intervallo e gli $m_{t,i}$, alla fine ; ed analogamente calcolare un valore per eccesso di S_t concentrando i $k_{t,i}$ alla fine di ciascun intervallo e gli $m_{t,i}$ al principio. Rappresentati poi questi valori su uno stesso sistema cartesiano di assi possiamo determinare per via grafica detta area, perequando eventualmente qualche valore eccezionale.

Analogo procedimento si può seguire per trovare l'area di probabile giacitura dei q_t ottenuti dagli S_t. Il carattere elementare di queste determinazioni non ci fa ritenere di dilungarci con l'applicazione al nostro materiale.

12. — Calcolo della probabilità di eliminazione entro un periodo poliennale. — Passiamo ora alla determinazione delle probabilità di eliminazione entro un periodo poliennale, dalle quali occorrerà poi distinguere le probabilità medie di eliminazione entro un anno delle persone di un gruppo poliennale di età.

Non avendo la funzione di sopravvivenza, e supposto di non avere tutti gli elementi per il calcolo dei saggi annuali, essendo alcuni di questi elementi dati per classi poliennali, indicando con $S_{t,t+n}$ il saggio medio costante per gli individui di età $t \vdash t+n$ di essere eliminati entro il periodo (di n anni), e con $m_{t,t+n}$ gli

eliminati nello stesso periodo, avremo:

$$S_{t,t+n} = \frac{m_{t,t+n}}{l_t \cdot n + \int_t^{t+n} (k_u - m_u)\,(t + n - u)\,du} \qquad (30)$$

nella quale per il calcolo dell'integrale al denominatore, sarà generalmente necessario ricorrere, al solito, a ipotesi particolari. La espressione approssimata più semplice si ottiene ammettendo l'uniforme distribuzione degli entrati e degli usciti come nella (19).

Da questo saggio si ricava poi la probabilità che un individuo di età precisa t sia eliminato entro il periodo $t \vdash t + n$,

$$/_nQ_t = 1 - e^{-n\ S_{t,t+n}} \qquad (31)$$

Questa probabilità potrebbe ricavarsi dalla espressione

$$/_nQ_t = \frac{m_{t,t+n}}{l_t + \int_t^{t+n} k_u \frac{q\,(u,\,t+n)}{/_nQ_t}\,du} \qquad (32)$$

ma l'ampiezza dell'intervallo $t \vdash t + n$ rende difficile l'assunzione di ipotesi accettabili per il quoziente $\frac{q\,(u,\,t+n)}{/_nQ_t}$, tenuto conto di quanto si è detto al *n.* 6.

Non mi pare che gli attuari abbiano mai adoperato queste formule per il calcolo delle probabilità in questione, poichè essi presuppongono di avere gli elementi per classi annuali, e, inoltre, nel caso delle probabilità di eliminazione per causa di morte, necessitando di approssimazioni molto buone, escludono di ricorrere a formule che possono portare a risultati non soddisfacenti.

Non altrettanto può dirsi, però, per altre cause di eliminazione che interessano particolarmente gli statistici.

Da un punto di vista teorico, il problema rientra nei casi precedentemente esaminati quando si suppone che il periodo anzichè uguale a 1 sia uguale ad n. Ma qualche chiarimento mi

pare utile per quei casi in cui si può procedere con metodi diversi.

Anzitutto, nel caso in cui ci siamo messi precedentemente, l'ipotesi di un saggio costante in tutto l'intervallo $t \vdash t + n$ è spesso poco plausibile (essendo generalmente sensibile l'influenza della età sui saggi di eliminazione [1]), ma, come abbiamo osservato nel n. 6, essa sarà da preferire, il più delle volte, all'ipotesi (4) che può essere inaccettabile.

Se, invece, abbiamo gli elementi per il calcolo dei saggi annuali, si avrà:

$$ {}_nQ_t = 1 - e^{-\int_t^{t+n} \mu_s \, d_s} = 1 - e^{-\left(\int_t^{t+1} + \cdots\cdots + \int_{t+n-1}^{t+n}\right)} \backsimeq \tag{33} $$

$$ \backsimeq 1 - e^{-\sum_{i=0}^{n-1} S_{t+i}} $$

che esprime la probabilità richiesta in funzione dei saggi medi annuali. Questa formula differisce dalla

$$ {}_nQ_t = 1 - p_t \cdot p_{t+1} \cdot \cdots\cdots \cdot p_{t+n-1} \tag{34} $$

quando le probabilità p_t sono ricavate in base a ipotesi diverse da quella dei saggi costanti.

Alle volte si procede in altro modo: si calcola prima

$$ S'_{t,t+n} = \frac{\sum_{i=0}^{n-1} m_{t+i}}{\sum_{i=0}^{n-1} M_{t+i}} \tag{35} $$

che è una media ponderata dei saggi annuali e poi si determina la corrispondente probabilità servendosi della (31), nella quale al posto di $S_{t,t+n}$ dato dalla (30) si sostituisce il valore $S'_{t,t+n}$.

Questo valore $S'_{t,t+n}$ coincide con $S_{t,t+n}$ dato dalla (30) qualora si potessero determinare le durate esatte di esposizione al

---

[1] Per i saggi di fecondità cfr. C. Gini, *Sulla curva della fecondità matrimoniale della donna secondo l'età*, XXII sessione dell'Istituto Internazionale di Statistica, Londra, 1934.

rischio di ciascun individuo. S'intende che ove non sia possibile calcolare le durate esatte di esposizione al rischio dei vari contingenti, possiamo avere diversi valori di $S'_{t,t+n}$, ciascuno in corrispondenza all'ipotesi scelta per il calcolo dei singoli M_{t+i} che figurano nella (35).

Il valore di $S'_{t,t+n}$ sostituito nella (31) al posto di $S_{t,t+n}$, permette ricavare ${}_nQ_t$, ma questo calcolo essendo sempre infirmato dall'ipotesi poco plausibile del saggio medio costante in tutto l'intervallo $(t, t+n)$, fa ritenere più esatto il valore che si ricava con la (33).

Agli effetti, dunque, della determinazione delle probabilità di eliminazione entro un periodo poliennale riteniamo più vantaggioso, generalmente, usare la media aritmetica semplice dei saggi corrispondenti ai singoli anni:

$$\frac{1}{n} \sum_{i=0}^{n-1} S_{t+i}$$

piuttosto che la media aritmetica ponderata $S'_{t,t+n}$ degli S_t.

Aggiungiamo, analogamente a quanto abbiamo detto alla fine del n. 10, che interpretando gli S_t come ordinate della funzione matematica che esprime μ_s, il calcolo di $\int_t^{t+n} \mu_s \, ds$, che figura nell'espressione esatta di ${}_nQ_t$, potrebbe eseguirsi ricorrendo a opportune formule di quadratura, facendo particolari ipotesi sull'andamento della funzione μ_s (ad es. quella semplice di un andamento parabolico).

Per esaminare in concreto le differenze che si riscontrano nei risultati ottenuti usando la (33), oppure la (31) insieme alla (35), oppure la (34) nell'ipotesi del metodo delle durate esatte per i p_t, applichiamo queste formule ai dati del Gini relativi alle coniugate della Case sovrane, mediatizzate e principesche, già richiamate [1]. Nella tabella seguente riportiamo pure i valori del saggio medio calcolato dal Gini con la (35).

Mi pare opportuno precisare che mentre il saggio medio costante si riferisce alla eliminazione per concepimento delle coniu-

---

[1] C. Gini, *Su la determinazione...*, cit., tav. II per gli S_t e tav. VIII per gli $S_{t,t+n}$.

gate del gruppo esaminato e di età indicata dalla classe scritta nella prima colonna della tabella, le ultime tre colonne conten-

TAV. VI.

Valori approssimati delle probabilità ${}_n/Q_t$ di eliminazione entro un periodo poliennale.

Età	Saggio medio $S'_{t,\ t+n}$ (35)	$-\Sigma S_{t+i}$ / I—e (33)	$-n S'_{t,\ t+n}$ / I—e (31) e (35)	Formula (34) nell'ipotesi (4) per i p_t
16–19	1,0711	0,9884	0,9862	(^)
20–24	0,6275	0,9663	0,9566	0,9729
25–29	0,3413	0,8296	0,7550	0,8337
20–34	0,1715	0,5790	0,5758	0,5804
35–39	0,0829	0,3381	0,3393	0,3401
40–44	0,0195	0,0902	0,0929	0,0902
45–46	0,0048	0,0095	0,0096	0,0095
16–46	0,1894	0,99998	0,9972	(^)

(^) Il primo valore e quello relativo all'intervallo totale non possono calcolarsi perchè la probabilità q_{16} calcolata nell'ipotesi (4) è superiore all'unità (Cfr. tav. X del lav. cit. in (1) del Gini).

gono le probabilità — approssimate — che una coniugata del gruppo esaminato, di età indicata dal limite inferiore della classe di età, sia eliminata per concepimento nel periodo quinquennale seguente, o nell'intervallo 16-46 per i dati dell'ultima riga. La differenza tra le probabilità corrispondenti, ottenute con la (33) (che riteniamo la più esatta) e la (31) con la (35) è particolarmente notevole in corrispondenza dell'età di 25 anni, sensibile a 20 anni, lieve alle altre età. I primi quattro valori calcolati con la (33) sono più grandi di quelli ottenuti con la (31) insieme alla (35), mentre il contrario avviene per i tre valori successivi. I valori che si ottengono, poi, dalla (34) nell'ipotesi (4) per i p_t differiscono sensibilmente in più da quelli ottenuti con la (33), ad eccezione degli ultimi due valori che sono identici. Non pos-

sono calcolarsi i ${}_nQ_t$ per $t = 16$ poichè, nella detta ipotesi, p_{16} è negativo. Molto elevata è la probabilità per una coniugata a 16 anni, del gruppo considerato, di concepire entro i primi quattro anni di matrimonio; per le stesse coniugate, la probabilità di concepire entro tutto il periodo di fecondità è quasi uguale all'unità.

13. — Calcolo della probabilità media di eliminazione di un gruppo avente età $t, t+n$. — Diversa dalla precedente è la probabilità media che una testa di età $t \vdash t+n$ sia eliminata entro un anno, per una data causa.

Indicando questa probabilità media col simbolo $Q_{t,t+n}$ essa sarà data dal rapporto tra gli eliminati $m_{t\,t+n}$ di età $t \vdash t+n$ e gli esposti $E_{t,t+n}$ della stessa età, cioè:

$$Q_{t,t+n} = \frac{m_{t\,t+n}}{E_{t,t+n}} = \frac{m_t + m_{t+1} + \ldots + m_{t+n-1}}{E_t + E_{t+n} + \ldots + E_{t+n-1}} \qquad (36)$$

dove $E_{t,t+n}$ è diverso da $M_{t,t+n} = \int_t^{t+n} E_s\, ds$ che figura e denominatore della (30) ed è anche diverso dal denominatore della (32). Il valore di $Q_{t,t+n}$ si ottiene, com'è noto, facendo una media ponderata di $q_t,\ q_{t+1} \ldots\ldots q_{t+n-1}$ con pesi uguali rispettivamente a $E_t,\ E_{t+1} \ldots\ldots E_{t+n-1}$.

Per $n = 1$ si ha: $Q_{t,t+1} = m_{t,t+1} \,/\, E_{t,t+1}$ che è teoricamente diversa da ${}_1Q_t = q_t = m_{t,t+1} \,/\, E_t = 1 - e^{-\int_t^{t+1} \mu_s\, ds}$ (che si riferisce agli esposti di età precisa t) anche se, in pratica, si assume q_t come esprimente la probabilità di eliminazione entro un anno degli individui di età $t - \frac{1}{2}$, $t + \frac{1}{2}$ o $t, t+1$ come per $Q_{t,t+1}$.

Per periodi unitari, quindi, o per gruppi unitari di individui (ad es. tutti gli individui di un certo stato civile, nel caso che si determinino le probabilità di eliminazione per stato civile), i due concetti in generale non si distinguono, mentre ciò è opportuno tutte le volte che la mancata distinzione può generare confusione.

Se si conoscesse la legge (incognita) di eliminazione si potrebbe determinare una età $t + \varepsilon$, compresa tra t e $t+n$, tale

da poter ritenere il gruppo di esposti di età indicata equivalente a un gruppo di esposti aventi tutti l'età $t + \varepsilon$. Ma la scelta di una legge approssimata diventa difficile se ci riferiamo a cause di eliminazione diverse da quella di morte. In via approssimata si potrebbe pensare che $Q_{t,t+n}$ esprima la probabilità di eliminazione entro un anno degli individui di età precisa $t + \frac{n}{2}$.

Rimanendo nella ipotesi di un saggio medio costante in ciascun intervallo che supponiamo di un anno, e supposto sempre che la collettività non formi gruppo chiuso, per ricavare gli E_{t+i} dagli S_{t+i} basta determinare prima $q_{t+i} = 1 - e^{-S_{t+i}}$ e poi gli E_{t+i} eseguendo il rapporto tra gli eliminati m_{t+n} e i q_{t+n}. Perciò

$$Q_{t,t+n} = \frac{\sum_{i=0}^{n-1} m_{t+i}}{\sum_{i=0}^{n-1} \frac{m_{t+i}}{q_{t+i}}} \simeq \frac{\sum_{i=0}^{n-1} m_{t+i}}{\sum_{i=0}^{n-1} \frac{m_{t+i}}{1 - e^{-S_{t+i}}}} \qquad (37)$$

mentre il valore esatto si avrebbe sostituendo, nell'esponente di e, la quantità: $\int_{t+i-1}^{t+i} \mu_s \, ds$ al posto di S_{t+i}.

Se questi saggi, costanti nei singoli anni, variano da anno ad anno, la precedente non si presta a semplificazioni. Se, invece, $S_t = S_{t+i} = \ldots\ldots = S_{t+n-1} = S_{t,t+n}$ dalla precedente relazione si ha, come era da aspettarsi:

$$Q_{t,t+n} \simeq 1 - e^{-S_{t,t+n}} \qquad (38)$$

diversa dalla (31) in quanto il periodo di esposizione è adesso di 1 anno.

Nella precedente relazione $S_{t,t+n}$ si calcola o con la (30) oppure, quando se ne hanno gli elementi, con la (35). Avendo però osservato che, almeno da un punto di vista logico, è poco plausibile l'ipotesi di un saggio costante nell'intervallo $t \vdash t + n$, ci pare più opportuno ricorrere, per il calcolo di $Q_{t,t+n}$ all'ultima della (37) che si fonda su ipotesi più accettabili. Questa espressione, poi, rispetto a quella che si ricava dalla (36) nell'ipotesi

(4) per il calcolo degli E_{t+i}, è da preferire per le ragioni più volte illustrate, tra cui, in particolare, quella detta nel n. 6.

Nella tavola VIII del lav. cit. in (1) del Gini è già riportato un confronto tra i $Q_{t,t+5}$ che si ricavano dalla (36) con gli E_{t+i} calcolati col metodo delle durate esatte e quelli ottenuti con la (38) insieme alla (35); i primi sono maggiori o uguali ai secondi, con differenze sensibili nei gruppi giovanili di spose. Valori assai prossimi ai secondi si ottengono applicando la (37). Nel prossimo paragrafo aggiungeremo un'altra applicazione.

14. — Applicazione delle precedenti probabilità. — Prima di chiudere questa nota, osserviamo che il passaggio dai saggi medi alle corrispondenti probabilità, pur richiedendo l'introduzione di una ipotesi, è molte volte opportuno, perchè, come abbiamo osservato nel n. 2, i saggi medi non ammettono un limite superiore costante, mentre le probabilità, variando costantemente tra 0 e 1, si prestano meglio ai confronti; inoltre ad esse sono applicabili immediatamente i teoremi del Calcolo delle Probabilità, i quali permettono di calcolare facilmente, dalle probabilità elementari, alcune probabilità più complesse.

Osserviamo poi che, molte volte, per uno studio sintetico della causa di eliminazione fatto per scopi statistici, è scomodo oltre che faticoso servirsi di parecchie decine di q_t. [1]. In tali casi, pur potendosi calcolare le probabilità annuali, è opportuno calcolare le probabilità poliennali (quinquennali o decennali), e, alle volte, per una sintesi maggiore, si considerano i quozienti medi generici i quali si ottengono riunendo in un unico gruppo gli individui di tutte le età. Parliamo, per semplicità di linguaggio, di età, ma quello che diciamo vale anche quando si hanno i quozienti di eliminazione distinti per altre caratteristiche (sesso, stato civile, professione, ecc.) e si vuole un quoziente generico.

Per un ulteriore esame concreto delle differenze che si riscontrano nei valori dei quozienti poliennali e in quelli generici

---

[1] C. Gini, *Quelques considérations au sujet de la construction des nombres indices des prix et des questions analogues*, in « Metron », Vol. IV, 1924. M. Boldrini, *Demografia*, Giuffré, 1946, pag. 124,

con i metodi esposti, prendiamo in esame i risultati del Paglino sulla fecondità matrimoniale della donna [1].

Il Paglino calcola prima il saggio generico di fecondità, poi i saggi quinquennali e poi quelli annuali, i quali costituirebbero la *tavola di fecondità* matrimoniale della donna, per la popolazione italiana 1930-32. Il Paglino, però, non passa dai saggi

TAV. VII.

Saggi e probabilità (grezze) di fecondità matrimoniale della donna italiana (1930-32).

Età	S_t	$q_t = 1 - e^{-S_t}$	Età	S_t	$q_t = 1 - e^{-S_t}$
15	0,3006	0,2596	33	0,2332	0,2080
16	0,2758	0,2410	34	0,2262	0,2024
17	0,3591	0,3017	35	0,2110	0,1902
18	0,5590	0,4282	36	0,2027	0,1835
19	0,7221	0,5143	37	0,1775	0,1625
20	0,8118	0,5559	38	0,1622	0,1497
21	0,6288	0,4668	39	0,1367	0,1278
22	0,5804	0,4403	40	0,1339	0,1253
23	0,5501	0,4231	41	0,0881	0,0843
24	0,5379	0,4160	42	0,0763	0,0735
25	0,4570	0,3668	43	0,0509	0,0496
26	0,4083	0,3352	44	0,0338	0,0332
27	0,3793	0,3157	45	0,0216	0,0214
28	0,3472	0,2933	46	0,0111	0,0110
29	0,2955	0,2558	47	0,0053	0,0053
30	0,3154	0,2705	48	0,0029	0,0029
31	0,2570	0,2266	49	0,0014	0,0014
32	0,2719	0,2381			

[1] F. PAGLINO, *Su la misura della fecondità matrimoniale della donna*, « Metron », Vol. XIV, nn. 2-3-4, 1941.

medi alle corrispondenti probabilità. Se fondiamo questo passaggio su una stessa ipotesi è chiaro che le conclusioni di questo autore non vengono affatto modificate ragionando sulle probabilità. Queste, però, permettono una più razionale valutazione delle differenze che si riscontrano tra i fenomeni a confronto (influenza delle diverse ipotesi a base dei calcoli, confronto ad es. tra la fecondità in Italia e in Francia).

Per i nostri scopi partiamo dai saggi annuali di fecondità del Paglino e ricaviamo i ${}_nQ_t$ e i $Q_{t\,t+n}$ sia per $n = 5$ sia per n uguale a tutto il periodo di fecondità.

Il Paglino ha determinato varie categorie di saggi annuali, corrispondenti ad altrettante ipotesi sulla eliminazione dalla collettività e sul loro rientro dopo il parto. Per brevità, fermiamo la nostra attenzione sui saggi indicati con $S_4(x)$ i quali (cfr. op. cit. pag. 74) darebbero la misura esatta della fecondità matrimoniale.

Nella tav. VII accanto ai saggi sono riportate le corrispondenti probabilità, ricavate come abbiamo detto, nell'ipotesi del saggio di eliminazione costante nell'intervallo di un anno.

Da questi quozienti annuali abbiamo ricavato quelli quinquennali e quello generico, relativo a tutte le età, secondo le varie formule indicate nella tavola VIII.

TAV. VIII.

Saggi e probabilità poliennali di fecondità matrimoniale della donna italiana (1930-32) *ottenuti con vari procedimenti*

Età	$S'_{t,\,t+5}$ (35)	$S_{t,\,t+5}$ (Cap. II del Paglino)	$Q_{t,\,t+5} = 1-e^{-S'_{t,\,t+5}}$	$Q_{t,\,t+5} = \frac{\Sigma m_{t+i}}{\Sigma \frac{m_{t+i}}{1-e^{-S_{t+i}}}}$ (37)	$Q_{t,\,t+5} = 1-e^{-S_{t,\,t+5}}$	${}_5Q_t = 1-e^{-\Sigma S_{t+i}}$
1	2	3	4	5	6	7
15–19	0,5827	0,7068	0,4416	0,4410	0,5068	0,8910
20–24	0,5840	0,5659	0,4223	0,4222	0,4322	0,9554
25–29	0,3720	0,3492	0,3107	0,3106	0,2948	0,8485
30–34	0,2599	0,2386	0,2289	0,2289	0,2123	0,7285
35–39	0,1781	0,1627	0,1631	0,1631	0,1502	0,5894
40–44	0,0763	0,0716	0,0735	0,0734	0,0691	0,3182
45–49	0,0087	0,0078	0,0087	0,0087	0,0078	0,0414
15–49	0,2063	0,1927	0,1864	0,1862	0,1751	0,99946

Il confronto tra i dati di questa tabella mostra subito le sensibili differenze tra gli $S'_{t,\,t+5}$, calcolati come media ponderata

dei saggi annuali, e gli $S_{t,t+5}$, che si ottengono con calcolo diretto, in base ad elementi poliennali. Le probabilità $Q_{t,t+5}$ delle colonne 4 e 6 risentono di queste differenze. Caratteristico è invece il fatto che i risultati della colonna 5 sono quasi uguali a quelli della colonna 4; il primo procedimento è consigliabile, quindi, quando si conoscono gli $S'_{t,t+5}$; se, invece, si hanno i q_t come nella tav. VII o nelle comuni tavole di eliminazione per morte, conviene servirsi della prima espressione indicata nella (37), il cui calcolo è rapido. Si ottengono così, senza dubbio, dei valori più esatti di quelli che si ottengono col calcolo diretto in base ai contingenti poliennali.

Molto diversi da $Q_{t,t+5}$ sono i valori di ${}_5Q_t$; ovviamente, la probabilità che una donna di età $t \vdash t + 5$ ha di concepire entro un anno è più piccola della probabilità che una donna di età t ha di concepire entro cinque anni. I primi sono più usati dagli statistici, i secondi dagli attuari. Agli effetti statistici basta evidentemente limitarsi a tre cifre decimali.

In particolare mentre è quasi uguale a uno la probabilità per una donna di 15 anni, s'intende del gruppo considerato, di concepire nel corso del suo periodo fecondo, la probabilità di concepimento di una donna di età 15-49 è appena 0,186.

Quest'ultimo valore e non il precedente può essere assunto come *probabilità generica* di concepimento, mentre il *saggio generico* di concepimento è il valore 0,206 (oppure 0,193 se procediamo con i contingenti quinquennali).

Partendo dai saggi quinquennali $S_{t,t+5}$ del Paglino e applicando la (37) si ha come probabilità generica il valore 0,174 un po' diverso dal valore 0,186 ricavato in base ai saggi annuali. Tale differenza può portare a risultati discordanti quando si procede a confronti tra collettività diverse, giacchè non possiamo stabilire a priori se l'approssimazione che si ottiene con gli elementi poliennali sia sempre per difetto (e in una certa misura) rispetto al suo valore più esatto ottenuto in base agli elementi annuali, oppure se l'approssimazione sia qualche volta per difetto e qualche volta per eccesso.

Le considerazioni di questi ultimi tre paragrafi, pur avendo carattere elementare, ci sono sembrate opportune in un lavoro di chiarificazione e di precisazione del campo di applicazione dei metodi per il calcolo dei vari quozienti di eliminazione.

15. — RIASSUNTO. — In questo lavoro abbiamo analizzato le ipotesi e i procedimenti che sono alla base della determinazione dei quozienti di eliminazione. In particolare ci siamo soffermati sul metodo delle durate esatte, che si fonda sull'ipotesi (4) messa in evidenza dal Cantelli, e sul metodo del saggio medio costante proposto dal Gini (in una forma completamente diversa da quella seguita dal Wittstein).

In relazione ad alcune obbiezioni mosse al metodo del Gini, abbiamo messo in evidenza:

1) l'ipotesi del saggio di eliminazione costante è di per sè sufficiente, senza aggiunta di altre ipotesi o di indispensabili procedimenti approssimati, a determinare la probabilità di eliminazione;

2) il fatto che la funzione integranda nella formula del Gini del saggio medio S_t sia $k_u - m_u$, mentre in quella che dà q_t nel metodo delle durate esatte sia soltanto k_u, non significa che il risultato dell'integrazione approssimata della prima funzione debba essere più errato di quello che si ottiene dall'integrazione approssimata della seconda; l'errore dipende sia dalla forma della funzione integranda sia dalla formula di quadratura che deve essere opportunamente scelta in base al reale o al presumibile andamento di detta funzione; è da tenere anche presente che per alcuni fenomeni — come quello di eliminazione per fecondità — le quantità $k_u - m_u$ sono inferiori, in valore assoluto, ai k_u;

3) che qualunque ipotesi semplificatrice sull'andamento di $k_u - m_u$ (nel caso che non si voglia fare direttamente uso di una formula di quadratura) serve a calcolare un valore approssimato del saggio medio di eliminazione, e, quindi, il discostarsi di una tale ipotesi semplificatrice dall'andamento che sarebbe coerente con il saggio di eliminazione costante, non costituisce un motivo di esclusione di una tale ipotesi, ma soltanto una causa di approssimazione: è la portata pratica dell'approssimazione che così si commette che deve far giudicare se una tale ipotesi semplificatrice sia da adottarsi o no per il calcolo. Del resto, lo stesso Cantelli riconosce che nelle applicazioni pratiche si può prescindere dal carattere teorico delle sue obbiezioni.

Nel n. 6 abbiamo poi messo in evidenza che l'ipotesi posta a base del metodo delle durate esatte è alle volte del tutto inaccettabile, come quando la causa di eliminazione porta a valori di q_t elevati: infatti, in tal caso, detta ipotesi implica una decrescenza notevolissima del saggio istantaneo di eliminazione tra t e $t+1$ che può essere molto in contrasto col reale andamento del saggio. Un tale contrasto non si verifica, invece, nell'ipotesi del Gini.

Pertanto, si può essere d'accordo con le preferenze del Cantelli quando si tratta di determinare le probabilità di morte per le età che interessano gli attuari, ma per molti altri fenomeni, nei quali la (4) è inaccettabile, il metodo del Gini risulta più appropriato.

Siccome ciascuno dei metodi esaminati per il calcolo delle probabilità di eliminazione fa ricorso a una ipotesi sull'andamento della legge incognita di eliminazione, ognuno di essi è soggetto ad errori sensibili se è adoperato in condizioni non rispondenti alle ipotesi su cui esso metodo è fondato.

Riteniamo, inoltre, che, se la rilevazione statistica dà gli elementi $k_{t,i}$ e $d_{t,i}$ indicati nelle (20), il procedimento da noi suggerito nel n. 10 sia un utile completamento del metodo del Gini, in quanto permette di calcolare i quozienti di eliminazione anche quando le condizioni del fenomeno facciano ritenere molto complicata la legge di variazione del saggio istantaneo e quindi non realistiche, in tutto l'intervallo $t \vdash t+1$, le ipotesi prese a base dai vari autori.

Esso permetterebbe anche di calcolare valori approssimati per difetto e valori approssimati per eccesso e quindi di determinare la striscia di probabile giacitura dei q_t.

Ci pare, infine, che le considerazioni svolte a proposito del calcolo delle probabilità di eliminazione entro un periodo poliennale e delle probabilità di eliminazione entro un anno per individui appartenenti a un gruppo poliennale di età, sia atto a chiarire due aspetti particolari di cui il primo è utile nel campo assicurativo, mentre il secondo serve molto nel campo statistico, in cui gli individui sono spesso raggruppati in classi per rendere più sintetici i confronti nel tempo e nello spazio.

KARL FREUDENBERG

Die Grenzen für die Anwendbarkeit des Gesetzes der kleinen Zahlen.

Wenn s Versuche mit der konstanten Wahrscheinlichkeit q für das Eintreffen des Ereignisses E gemacht werden, so ergibt sich nach dem Theorem von Bernoulli die Wahrscheinlichkeit P_m dafür, dass E gerade in m Fällen einträte, mit

$$P_{q,m} = \binom{s}{m} q^m (1-q)^{s-m}. \tag{1}$$

Betrachtet man nun s und m als grosse Zahlen, so kommt man in bekannter Weise zu der Exponentialfunktion von *Gauss.* Nimmt man dagegen an, dass s eine grosse Zahl sei, q aber so klein, dass $sq = \lambda$ immer noch eine ziemlich kleine Zahl bleibt, dann erhält man in noch einfacherer Weise den Grenzübergang, den *Poisson* zuerst ausgeführt hat (1), nämlich

$$P_{\lambda,m} = e^{-\lambda} \frac{\lambda^m}{m!}. \tag{2}$$

Diese Formel (2) hat v. Bortkiewicz unter dem Namen *Gesetz der kelinen Zahlen* in die Statistik eingeführt (2).

Da diese Formel unter der Voraussetzung abgeleitet ist, dass $s \to \infty$, sei nun nachgeprüft, bis zu welchen Werten von s herab sie noch anwendbar ist, wenn man irgendeinen Wert von

(1) S.D. Poisson, *Recherches sur la probabilité des jugements*, 1837.
(2) L. v. Bortkiewicz: *Das Gesetz der kleinen Zahlen.* Leipzig, 1898.

λ festhält ; es ist zweckmässiger, hier nicht von einem festen q, sondern von einem festen $s\,q = \lambda$ auszugehen, da dieses für die Verteilung wesentlich charakteristischer ist, wie schon daraus zu ersehen ist, dass in (2) nicht s und q einzeln auftreten, sondern nur ihr Produkt λ, d.h. der Erwartungswert für die Zahl der E in einer Reihe von s Beobachtungen.

Charlier hat eine exakte Formel für $P_{\lambda,m}$ entwickelt, die bei Darmois wiedergegeben ist [3]. Diese Formel stellt eine unendliche Reihe dar, deren Anwendung recht mühsam ist.

Hier soll nun für den genannten Zweck eine einfache Formel abgeleitet werden, die nicht nur geeignet ist, die Anwendbarkeit von (2) zu prüfen, sondern auch auf bequeme Weise den Grad der Näherung beträchtlich zu erhöhen.

Hierzu sei zunächst eine Funktion $V_{\lambda,m}$ eingeführt, die das Verhältnis des genauen Wertes (1) von $P_{\lambda,m}$ zum Näherungwerte (2) darstellt, also, wenn auch in (1) für q gesetzt wird $\frac{\lambda}{s}$:

$$V_{\lambda,m} = \frac{\binom{s}{m}\left(\frac{\lambda}{s}\right)^m\left(1-\frac{\lambda}{s}\right)^{s-m}}{e^{-\lambda}\,\frac{\lambda^m}{m!}}. \qquad (3)$$

Nach Kürzung ergibt dies

$$V_{\lambda,m} = \frac{s\,(s-1)\,(s-2)\ldots.(s-m+1)\left(1-\frac{\lambda}{s}\right)^{s-m}}{s^m\,e^{-\lambda}}. \qquad (4)$$

Setzt man nun die Gleichung an

$$\left(1-\frac{\lambda}{s}\right)^{s-m} = e^{-x}, \qquad (5)$$

so folgt beim Logarithmieren

$$x = (s-m)\left(\frac{\lambda}{s}+\frac{\lambda^2}{2\,s^2}+\frac{\lambda^3}{3\,s^3}+\ldots\right),$$

---

[3] G. Darmois : *Statistique mathématique*. Paris 1928, p. 84 ss.

oder, da s eine einigermassen grosse Zahl sein soll, unter Vernachlässigung der Glieder mit zweiten und höheren Potenzen von s im Nenner:

$$x = \frac{\lambda}{2s}(2s + \lambda - 2m)$$

Setzt man diesen Wert von x in (5) ein und ersetzt dann in (4) das $\left(1 - \frac{\lambda}{s}\right)^{s-m}$ durch den so gewonnenen Ausdruck, so ergibt sich:

$$V_{\lambda,m} = \frac{s\,(s-1)\,(s-2)\ldots(s-m+1)}{s^m}\, e^{\frac{2m\lambda-\lambda^2}{2s}}. \qquad (6)$$

Da der Exponent von e eine kleine Zahl ist, kann man die Potenz als Reihe entwickeln und hierbei wieder die Glieder mit zweiten und höheren Potenzen von s im Nenner vernachlässigen, so dass also

$$V_{\lambda,m} = \frac{s\,(s-1)\,(s-2)\ldots(s-m+1)}{s^m}\left(1 - \frac{\lambda^2 - 2m\lambda}{2s}\right). \qquad (7)$$

Fasst man nun im Zähler des ersten Bruches je zwei Glieder zusammen, die gleichweit von den Enden abstehen, und ersetzt jedes solche Paar durch den fast genau gleich grossen Wert $s\,(s-m+1)$, (bei ungeradem m gilt für das übrig bleibende Mittelglied das Entsprechende), so wird

$$V_{\lambda,m} = \frac{[s\,(s-m+1)]^{\frac{m}{2}}}{s^m}\left(1 - \frac{\lambda^2 - 2m\lambda}{2s}\right) =$$

$$= \left(\frac{s-m+1}{s}\right)^{\frac{m}{2}}\left(1 - \frac{\lambda^2 - 2m\lambda}{2s}\right),$$

somit

$$V_{\lambda,m} = \left(1 - \frac{m+1}{s}\right)^{\frac{m}{2}}\left(1 - \frac{\lambda^2 - 2m\lambda}{2s}\right). \qquad (8)$$

Da $\frac{m+1}{s}$ gegen 1 klein ist, kann die Potenz im ersten Gliede nach der Binomialformel entwickelt werden, und mit Vernachlässigung der Glieder höherer Ordnung ergibt dies

$$V_{\lambda,m} = \left[1 - \frac{m\,(m-1)}{2\,s}\right]\left(1 - \frac{\lambda^2 - 2\,m\,\lambda}{2\,s}\right),$$

und abermals unter gleicher Vernachlässigung

$$V_{\lambda,m} = 1 - \frac{m\,(m-1)}{2\,s} - \frac{\lambda^2 - 2\,m\,\lambda}{2\,s},$$

d.h.

$$V_{\lambda,m} = 1 - \frac{(m-\lambda)^2 - m}{2\,s}. \qquad (9)$$

Diese Formel (9), die für die weitere Untersuchung grundlegend ist, zeigt zunächst, dass die Abweichung des $V_{\lambda,m}$ von 1 mit der ersten Potenz von s abnimmt. Dann zeigt sie, dass bei gegebenem s und λ das Verhältnis $V_{\lambda,m}$ von 1 als parabolische Funktion abweicht, wobei folgende Werte für bestimmte m hervorgehoben seien:

$$V_{\lambda,0} = 1 - \frac{\lambda^2}{2\,s}$$

$$V_{\lambda,1} = 1 - \frac{\lambda^2 - 2\,\lambda}{2\,s}$$

$$V_{\lambda,\lambda} = 1 + \frac{\lambda}{2\,s}$$

$$V_{\lambda,\lambda+1} = 1 + \frac{\lambda}{2\,s}$$

$$V_{\lambda,2\lambda} = 1 - \frac{\lambda^2 - 2\,\lambda}{2\,s}$$

$$V_{\lambda,2\lambda+1} = 1 - \frac{\lambda^2}{2\,s}$$

Der Gipfel von $V_{\lambda,m}$ liegt bei $m = \lambda + \frac{1}{2}$, und die Werte liegen über 1, solange m von $\lambda + \frac{1}{2}$ um einen geringeren absoluten Betrag als $\frac{\sqrt{4\lambda+1}}{2}$ abweicht. Fällt m ausserhalb dieser Grenzen, so wird $V_{\lambda,m} < 1$; bei $m = \lambda + \frac{1}{2} \pm \frac{\sqrt{4\lambda+1}}{2}$ ist $V_{\lambda,m} = 1$. Auf jeden Fall ist $V_{\lambda,m} > 1$ für $m = \lambda$ und $m = \lambda + 1$ bzw. bei nicht ganzzahligem λ für die zwischen λ und $\lambda + 1$ fallende ganze Zahl, und es ist auf jeden Fall < 1 für $m = 0$ und für alle $m \geqq 2\lambda + 1$.

Mittels Formel (9) kann man nun dazu übergehen, die Näherung ersten Grades, die durch die Formel von Poisson (2) gegeben ist, in bequemer Weise durch eine Näherung zweiten Grades zu ersetzen. Der Kürze halber sollen im folgenden immer die nachstehenden Bezeichnungen gebraucht werden:

$A_{\lambda,m}$ für die Näherung ersten Grades nach Poisson (2);

$B_{\lambda,m}$ für die nunmehr darzulegende Näherung zweiten Grades;

$C_{\lambda,m}$ für die exakte Formel nach Bernoulli (1).

Nach der Definition (3) ist $V_{\lambda,m}$ näherungsweise das Verhältnis zwischen $C_{\lambda,m}$ und $A_{\lambda,m}$, d.h. $V_{\lambda,m} = \frac{B_{\lambda,m}}{A_{\lambda,m}}$, also $B_{\lambda,m} = A_{\lambda,m} V_{\lambda,m}$, daher wegen (9):

$$B_{\lambda,m} = e^{-\lambda} \frac{\lambda^m}{m!} \left[1 - \frac{(m-\lambda)^2 - m}{2s} \right]. \tag{10}$$

Die Verwendung von $B_{\lambda,m}$ wird dadurch sehr erleichtert, dass analog wie für $A_{\lambda,m}$ und $C_{\lambda,m}$ auch für diesen Ausdruck exakt gilt:

$$\sum_0^\infty m B_{\lambda,m} = 1. \tag{11}$$

Es ist nämlich, wie man leicht erkennt,

$$\sum_0^\infty m \frac{\lambda^m m}{m!} = \lambda e^\lambda \qquad \text{und} \qquad \sum_0^\infty m \frac{\lambda^m m^2}{m!} = \lambda e^\lambda + \lambda^2 e^\lambda$$

so dass

$$\sum_{o}^{\infty} m \frac{e^{-\lambda}\lambda^m}{m!} - \frac{e^{-\lambda}}{2s}\left(\sum_{o}^{\infty} m \frac{\lambda^m m^2}{m!} - 2\lambda \sum_{o}^{\infty} m \frac{\lambda^m m}{m!} + \right.$$

$$\left. + \lambda^2 \sum_{o}^{\infty} m \frac{\lambda^m}{m!} - \sum_{o}^{\infty} m \frac{\lambda^m m}{m!}\right) = \sum_{o}^{\infty} m \frac{e^{-\lambda}\lambda^m}{m!} -$$

$$- \frac{e^{-\lambda}}{2s}(\lambda e^{\lambda} + \lambda^2 e^{\lambda} - 2\lambda^2 e^{\lambda} + \lambda^2 e^{\lambda} - \lambda e^{\lambda}) = \sum_{o}^{\infty} m \frac{e^{-\lambda}\lambda^m}{m!} = 1$$

qu.e.d.

Allerdings gilt für die Summation der $B_{\lambda,m}$ (aber ebenso auch für die der $A_{\lambda,m}$!) der Einwand, dass diese Summation nicht über $m = s$ hinausgehen könne und dass innerhalb dieser Grenze die Summe über m nicht genau 1 ergebe ; doch ist dies völlig bedeutungslos, wenn man nur ein einigermassen vernünftiges Verhältnis von λ und s annimmt, was durch das späterhin folgende Zahlenbeispiel verdeutlicht werden wird.

Ferner können die $B_{\lambda,m}$ (dies im Gegensatze zu den $A_{\lambda,m}$) für manche Werte von m sogar negativ ausfallen, wie sich aus Formel (10) ergibt, und zwar kann dies wegen des parabolischen Verlaufes der Funktion $V_{\lambda,m}$ gemäss Formel (9) sowohl für sehr kleine als auch für sehr grosse m in Betracht kommen. Was nun die erstgenannten betrifft, so wird $B_{\lambda,m}$ für $m = 0$ negativ, wenn $s < \frac{\lambda^2}{2}$; für ein so kleines s darf also die Formel nicht angewandt werden ; ist $m > 0$, aber unterhalb $\lambda + \frac{1}{2}$ so könnte $B_{\lambda,m}$ nur für noch kleinere s negativ werden. Für grosse m sieht man andererseits aus (10), dass $B_{\lambda,m}$ dann negativ würde, wenn

$$m > \lambda + \frac{1}{2} + \sqrt{\left(\lambda + \frac{1}{2}\right)^2 + 2s - \lambda^2} =$$

$$= \lambda + \frac{1}{2} + \sqrt{\lambda + \frac{1}{4} + 2s} .$$

Da nun gemäss dem Vorstehenden die Mindestforderung $s > \frac{\lambda^2}{2}$ gelten muss, kann unter Berücksichtigung derselben $B_{\lambda,m}$ im absteigenden Aste seines Verlaufes erst negativ werden, wenn $m > \lambda + \frac{1}{2} + \sqrt{\lambda + \frac{1}{4} + \lambda^2} = 2\lambda + 1$. Wendet man nun auf $m = 2\lambda$ die Formel von Stirling an und schreitet von da zu $m = 2\lambda + 1$ fort, so findet man für dieses m:

$$e^{-\lambda} \frac{\lambda^m}{m!} = \left(\frac{e}{4}\right)^{\lambda} \frac{\lambda}{2(2\lambda+1)\sqrt{\pi\lambda}\left(1 + \frac{1}{24\lambda} + \ldots\right)}. \qquad (12)$$

Wenn der Wert von λ nicht klein ist, so wird dieser Ausdruck so winzig, dass es für die Gesamtverteilung bedeutungslos ist, wenn für diese grossen Werte von m die von $B_{\lambda,m}$ negativ werden. Ist dagegen λ klein, so erkennt man aus (12), dass dann die Mindestforderung $s > \frac{\lambda^2}{2}$ unzureichend wird. Ein weit brauchbareres Kriterium für die notwendige Mindestgrösse von s bei gegebenem λ wird sich ohnehin im weiteren Verlaufe dieser Untersuchung ergeben.

Zunächst sei an einigen Beispielen die vollständige Verteilung der Wahrscheinlichkeiten für das m-malige Eintreten von E bei verhältnismässig kleinen s dargetan, um zu zeigen, wie weit dann einerseits die Werte nach der ersten Näherung, d.h. die $A_{\lambda,m}$ und andererseits die nach der zweiten Näherung, d.h. die $B_{\lambda,m}$ von den exakten Werten, d.h. den $C_{\lambda,m}$ abweichen. Hierfür sind als Werte von s die Zahlen 100 und 50 gewählt, als λ sechs Zahlen, die von 0,5 bis 4 ansteigen. Die Berechnungen sind auf neun Dezimalen ausgeführt, alle Zahlen innerhalb der Tabelle in Einheiten der neunten Dezimale ausgedrückt.

Die Zahlen dieser Tabelle zeigen vorläufig rein empirisch, dass die Differenzen $|B_{\lambda,m} - C_{\lambda,m}|$ im Vergleiche zu denen $|A_{\lambda,m} - C_{\lambda,m}|$ sehr klein sind. Ausgenommen sind nur zwei Fälle: Der erste, dass $B_{\lambda,m} = A_{\lambda,m}$, was nur möglich ist, wenn $(m-\lambda)^2 - m = 0$, d.h. wenn λ so ist, dass $\lambda + \frac{1}{2} \pm \frac{\sqrt{4\lambda+1}}{2}$ eine ganze Zahl ergibt wie z.B. bei $\lambda = 2$ (oder angenähert, wenn dieser Ausdruck einer ganzen Zahl nahe kommt); der zweite, dass die m gegenü-

λ	m	$10^9 A_{\lambda,m}$	$s = 100$				$s = 50$			
			$10^9 B_{\lambda,m}$	$10^9 C_{\lambda,m}$	$10^9(A_{\lambda,m} - C_{\lambda,m})$	$10^9(B_{\lambda,m} - C_{\lambda,m})$	$10^9 B_{\lambda,m}$	$10^9 C_{\lambda,m}$	$10^9(A_{\lambda,m} - C_{\lambda,m})$	$10^9(B_{\lambda,m} - C_{\lambda,m})$
0,5	0	606 530 660	605 772 497	605 770 436	760 224	2 061	605 014 333	605 006 067	1 524 593	8 266
	1	303 265 330	304 402 575	304 407 254	— 1 141 924	— 4 679	305 539 820	305 558 620	— 2 293 290	— 18 800
	2	75 816 333	75 721 563	75 719 392	96 941	2 171	75 626 792	75 618 042	198 291	8 750
	3	12 636 055	12 430 719	12 429 649	206 406	1 070	12 225 383	12 221 098	414 957	4 285
	4	1 579 507	1 514 352	1 514 668	64 839	— 316	1 449 198	1 450 484	129 023	— 1 286
	5	157 951	145 907	146 139	11 812	— 232	133 863	134 792	23 159	— 929
	6	13 163	11 567	11 627	1 536	— 60	9 971	10 212	2 951	— 241
	7	940	774	785	155	— 11	609	648	292	— 39
	> 7	61	46	50	11	— 4	31	37	24	— 6
1	0	367 879 441	366 040 044	366 032 341	1 847 100	7 703	364 200 647	364 169 680	3 709 761	30 967
	1	367 879 441	369 718 838	369 729 638	— 1 850 197	— 10 800	371 558 235	371 601 714	— 3 722 273	— 43 479
	2	183 939 721	184 859 420	184 864 819	— 925 098	— 5 399	185 779 118	185 800 857	— 1 861 136	— 21 739
	3	61 313 240	61 006 674	60 999 166	314 074	7 508	60 700 108	60 669 668	643 572	30 440
	4	15 328 310	14 945 102	14 941 715	386 595	3 387	14 561 895	14 548 339	779 971	13 556
	5	3 065 662	2 897 051	2 897 787	167 875	— 736	2 728 439	2 731 525	334 137	— 3 086
	6	510 944	462 404	463 451	47 493	— 1 047	413 865	418 091	92 853	— 4 226
	7	72 992	62 408	62 863	10 129	— 455	51 824	53 632	19 360	— 1 808
	8	9 124	7 254	7 382	1 742	— 128	5 383	5 883	3 241	— 500
	9	1 014	735	762	252	— 27	456	560	454	— 104
	10	101	65	70	31	— 5	29	47	54	— 18
	> 10	10	5	6	4	— 1	1	4	6	— 3

λ	m	$10^9 A_{\lambda,m}$	$s = 100$				$s = 50$			
			$10^9 B_{\lambda,m}$	$10^9 C_{\lambda,m}$	$10^9(A_{\lambda,m} - C_{\lambda,m})$	$10^9(B_{\lambda,m} - C_{\lambda,m})$	$10^9 B_{\lambda,m}$	$10^9 C_{\lambda,m}$	$10^{11}(A_{\lambda,m} - C_{\lambda,m})$	$10^{11}(B_{\lambda,m} - C_{\lambda,m})$
1,5	0	223 130 160	220 619 946	220 608 910	2 521 250	11 036	218 109 731	218 065 375	5 064 785	44 356
	1	334 695 240	335 950 347	335 952 656	— 1 257 416	— 2 309	337 205 454	337 214 498	— 2 519 258	— 9 044
	2	251 021 430	253 217 868	253 243 499	— 2 222 069	— 25 631	255 414 305	255 518 202	— 4 496 772	—103 897
	3	125 510 715	125 981 380	125 979 000	— 468 285	2 380	126 452 045	126 441 997	— 931 282	10 048
	4	47 066 518	46 537 020	46 522 702	543 816	14 318	46 007 521	45 949 282	1 117 236	58 239
	5	14 119 955	13 608 107	13 602 577	517 378	5 530	13 096 258	13 074 229	1 045 726	22 029
	6	3 529 989	3 278 477	3 279 809	250 180	— 1 332	3 026 966	3 032 682	497 307	— 5 716
	7	756 426	668 491	670 708	85 718	— 2 217	580 557	589 564	166 862	— 9 007
	8	141 830	117 542	118 736	23 094	— 1 194	93 253	98 007	43 823	— 4 754
	9	23 638	18 054	18 483	5 155	— 429	12 469	14 145	9 493	— 1 676
	10	3 546	2 442	2 561	985	— 119	1 339	1 794	1 752	— 455
	> 10	553	326	359	194	— 33	102	225	328	— 123
2	0	135 335 283	132 628 577	132 619 556	2 715 727	9 021	129 921 872	129 885 794	5 449 489	36 078
	1	270 670 566	270 670 566	270 652 155	18 411	18 411	270 670 566	270 595 403	75 163	75 163
	2	270 670 566	273 377 272	273 413 912	— 2 743 346	— 36 640	276 083 977	276 232 807	— 5 562 241	—148 830
	3	180 447 044	182 251 514	182 275 941	— 1 828 897	— 24 427	184 055 985	184 155 205	— 3 708 161	— 99 220
	4	90 223 522	90 223 522	90 207 991	15 531	15 531	90 223 522	90 159 319	64 203	64 203
	5	36 089 409	35 367 621	35 346 805	742 604	20 816	34 645 833	34 561 072	1 528 337	84 761
	6	12 029 803	11 428 313	11 421 587	608 216	6 726	10 826 823	10 800 335	1 229 468	26 488
	7	3 437 087	3 127 749	3 130 114	306 973	— 2 365	2 818 411	2 828 659	608 428	— 10 248
	8	859 272	738 974	742 604	116 668	— 3 630	618 676	633 502	225 770	— 14 826
	9	190 949	152 759	154 920	36 029	— 2 161	114 569	123 181	67 768	— 8 612
	10	38 190	27 879	28 771	9 419	— 892	17 567	21 045	17 145	— 3 478
	11	6 944	4 514	4 804	2 140	— 290	2 083	3 188	3 756	— 1 105
	> 11	1 365	740	840	525	— 100	116	490	875	— 374

Segue: Tab.

λ	m	$10^9 A_{\lambda,m}$	$s = 100$				$s = 50$			
			$10^9 B_{\lambda,}$	$10^9 C_{\lambda,m}$	$10^9(A_{\lambda,m} - C_{\lambda,m})$	$10^9(B_{\lambda,m} - C_{\lambda,m})$	$10^9 B_{\lambda:m}$	$10^9 C_{\lambda,m}$	$10^9(A_{\lambda,m} - C_{\lambda,m})$	$10^9(B_{\lambda,m} - C_{\lambda,m})$
3	0	49 787 068	47 546 650	47 552 908	2 234 160	— 6 258	45 306 232	45 330 227	4 456 841	— 23 995
	1	149 361 205	147 120 787	147 069 612	2 291 593	51 175	144 880 369	144 672 532	4 688 673	207 837
	2	224 041 808	225 162 017	225 152 963	— 1 111 155	9 054	226 282 226	226 243 170	— 2 201 362	39 056
	3	224 041 808	227 402 435	227 474 127	— 3 432 319	— 71 692	230 763 062	231 056 900	— 7 015 092	—293 838
	4	168 031 356	170 551 826	170 605 596	— 2 574 240	— 53 770	173 072 297	173 292 675	— 5 261 319	—220 378
	5	100 818 813	101 322 907	101 308 065	— 489 252	14 842	101 827 001	101 763 358	— 944 545	63 643
	6	50 409 407	49 653 266	49 609 619	799 788	43 647	48 897 125	48 716 501	1 692 906	180 624
	7	21 604 031	20 631 850	20 603 701	1 000 330	28 149	19 659 668	19 545 830	2 058 201	113 838
	8	8 101 512	7 412 883	7 407 774	693 738	5 109	6 724 255	6 705 883	1 395 629	18 372
	9	2 700 504	2 335 936	2 341 977	358 527	— 6 041	1 971 368	1 997 497	703 007	— 26 129
	10	810 151	652 172	659 134	151 017	— 6 962	494 192	522 749	287 402	— 28 557
	11	220 950	162 398	166 791	54 159	— 4 393	103 847	121 334	99 616	— 17 487
	12	55 238	36 181	38 259	16 979	— 2 078	17 124	25 170	30 068	— 8 046
	≥ 12	16 149	8 692	9 474	6 675	— 782	1 234	6 174	9 975	— 4 940
4	0	18 315 639	16 850 388	16 870 319	1 445 320	— 19 931	15 385 137	15 466 476	2 849 163	— 81 339
	1	73 262 556	70 332 054	70 292 997	2 969 559	39 057	67 401 552	67 245 547	6 017 009	156 005
	2	146 525 111	145 059 860	144 979 307	1 545 804	80 553	143 594 608	143 262 253	3 262 858	332 355
	3	195 366 815	197 320 483	197 332 946	— 1 966 131	— 12 463	199 274 151	199 321 395	— 3 954 580	— 47 244
	4	195 366 815	199 274 151	199 388 497	— 4 021 682	114 346	203 181 488	203 054 400	— 8 287 054	—472 981
	5	156 293 452	159 419 321	159 510 798	— 3 217 346	— 91 477	162 545 190	162 923 575	— 6 630 123	—378 385
	6	104 195 635	105 237 591	105 232 818	— 1 037 183	4 773	106 279 548	106 254 505	— 2 058 870	25 043
	7	59 540 363	58 944 959	58 880 267	660 096	64 692	58 349 556	58 076 997	1 403 366	272 559
	8	29 770 181	28 579 374	28 520 129	1 250 052	59 245	27 388 567	27 144 083	2 625 498	243 884
	9	13 231 192	12 172 697	12 147 463	1 083 729	25 234	11 114 202	11 015 234	2 215 958	98 968
	10	5 292 477	4 604 455	4 605 913	686 564	— 1 458	3 916 433	3 927 170	1 365 307	— 10 737
	11	1 924 537	1 558 875	1 570 198	354 339	— 11 323	1 193 213	1 241 793	682 744	— 48 580
	12	641 512	474 719	485 235	156 277	— 10 516	307 926	350 941	290 571	— 43 015
	13	197 388	130 276	137 033	60 355	— 6 757	63 164	89 203	108 185	— 26 039
	≥ 13	76 327	40 797	46 080	30 247	— 5 283	5 265	25 759	50 568	— 20 494

ber dem betreffenden λ extrem hoch sind. In diesen zwei Fällen liegen aber die absoluten Werte der beiden betrachteten Differenzen so niedrig, dass sie bei der Prüfung des Näherungsgrades der Formeln $A_{\lambda,m}$ und $B_{\lambda,m}$ unberücksichtigt bleiben können.

Nunmehr kann die Prüfung der Hauptfrage bogonnen werden, nämlich wie gross bei gegebenem s und λ der absolute Betrag der Differenzen $A_{\lambda,m} - C_{\lambda,m}$ und $B_{\lambda,m} - C_{\lambda,m}$ für irgendein m im Höchstfalle werden kann.

Dies sei zunächst für $A_{\lambda,m} - C_{\lambda,m}$ untersucht. Indem man vorläufig von der empirisch festgestellten Tatsache ausgeht, die später auch algebraisch untersucht werden soll dass $|B_{\lambda,m} - C_{\lambda,m}|$ gegenüber $|A_{\lambda,m} - C_{\lambda,m}|$ unbedeutend ist, kann vorläufig diese letztgenannte Differenz durch die verhältnismässig wenig hiervon abweichende Differenz $|A_{\lambda,m} - B_{\lambda,m}|$ ersetzt werden, Wie sich aus (10) ergibt, ist

$$A_{\lambda,m} - B_{\lambda,m} = e^{-\lambda} \frac{\lambda^m}{m!} \frac{(m-\lambda)^2 - m}{2s} . \tag{13}$$

Dieser Ausdruck sei mit f_m bezeichnet. Dann ist also zu ermitteln, für welche Werte von m Extremwerte von f_m auftreten. Differenziert man zu diesem Zwecke f_m nach m, nachdem $m!$ durch die Formel von Stirling (ohne Korrekturglieder) ausgedrückt ist, so erhält man:

$$\frac{df_m}{dm} = \frac{e^{-\lambda}}{2s\sqrt{2\pi m}} \left(\frac{\lambda e}{m}\right)^m \left[\left(m^2 - 2\lambda m + \lambda^2 - m\right) \left(ln\,\lambda - \right.\right.$$

$$\left.\left. - ln\,m - \frac{1}{2m} \right) + 2m - 2\lambda - 1 \right]. \tag{14}$$

Da $s > 0$, $\lambda > 0$ und $m \geqq 0$, muss der ganze Ausdruck, der vor der eckigen Klammer steht, positiv sein. Daher genügt es zur Bestimmung der Extremwerte von f_m, den Ausdruck innerhalb der eckigen Klammern $= 0$ zu setzen, also

$$\left(m^2 - 2\lambda m + \lambda^2 - m\right) \left(ln\,\lambda - ln\,m - \frac{1}{2m} \right) + 2m - 2\lambda - 1 = 0. \tag{15}$$

Diese transzendente Gleichung lässt sich nicht in geschlossener Form auflösen. Man sieht nur, dass sowohl für m=O als auch für $m=\infty$ die linke Seite $=-\infty$ wird und dass es daher für jeden Wert von λ eine gerade Anzahl von Auflösungen geben muss. Eine zahlenmässige Prüfung mit Näherungsmethode für die praktisch in Betracht kommenden Werte von λ ergibt, dass die Zahl der Auflösungen von λ abhängt und bei kleinen Werten desselben $=2$, bei grösseren $=4$ ist. Aus der Tabelle lässt sich der Sachverhalt nicht immer erkennnen, weil manche Werte von m, die sich so ergeben, nicht in die Nähe ganzer Zahlen fallen. Andererseits versagt die Gleichung (14) für m=O, wo auch die Formel von Stirling nicht brauchbar ist, und dieser Fall muss nachher besonders untersucht werden.

Jedenfalls also hat die Gleichung (15) mehrere Lösungen, d.h. der Ausdruck (13) hat mehrere Extremwerte, und es lässt sich nicht allgemein ermitteln, welcher von diesen die absolut grösste Abweichung von O aufweist. Es liegt nur nahe, ohne zwingend bewiesen werden zu können, dass diese grösste Abweichung in der Nähe von $m=\lambda$ liegen dürfte. Denn von den beiden Ausdrücken auf der rechten Seite von (13) hat $e^{-\lambda}\frac{\lambda^m}{m!}$ das Maximum für ein m im Bereiche zwischen $\lambda-1$ und λ, und $\frac{(m-\lambda)^2-m}{2s}$ hat den einen seiner Extremwerte bei $m=\lambda+\frac{1}{2}$. Tatsächlich nimmt der Ausdruck auf der linken Seite von (15) für $m=\lambda$ den Wert $-\frac{1}{2}$ an und sein Differentialquotient nach m an der gleichen Stelle den Wert $+3$, so dass die betreffende Auflösung von (15) näherungsweise $m=\lambda+\frac{1}{6}$ ergibt. Sofern dies keine ganze Zahl ist, muss die Auflösung einer nahe gelegenen ganzen Zahl entsprechen. Zur wesentlichen Vereinfachung der Rechnung wird nun vorläufig vorausgesetzt, λ sei eine ganze Zahl, so dass der betreffende Extremwert von (13) bei $m=\lambda$ liegt.

Die Tabelle bestätigt dann empirisch, dass für $m=\lambda$ die Differenz $|A_{\lambda,m}-B_{\lambda,m}|$ bei jedem λ grösser ist als für ein anderes m; ausgenommen kann der Fall m=O sein, dessen besondere Betrachtung im vorstehenden noch vorbehalten wurde.

Bezeichnet man den absoluten Betrag von $A_{\lambda,m} - B_{\lambda,m}$ mit $F_{\lambda,m}$, so zeigt Formel (13), dass diese Beträge sich bei $m = 0$ und $m = \lambda$ folgendermassen verhalten:

$$F_{\lambda,0} : F_{\lambda,\lambda} = e^{-\lambda}\lambda^2 : e^{-\lambda}\frac{\lambda^{\lambda+1}}{\lambda!}. \qquad (16)$$

Ersetzt man nun wieder $\lambda!$ durch seinen Wert nach *Stirling* (ohne Korrekturglieder), so wird dieses Verhältnis $= \lambda\sqrt{\lambda}\,e^{-\lambda}\sqrt{2\pi}$. Dieser Ausdruck, nach λ differenziert, ergibt $\sqrt{2\pi}\,e^{-\lambda}\sqrt{\lambda}\left(\frac{3}{2} - \lambda\right)$. Dies wird $= 0$ bei $\lambda = 0$, $\lambda = \frac{3}{2}$ und $\lambda = \infty$, doch zeigt die nochmalige Differentiation, dass nur bei $\lambda = \frac{3}{2}$ ein Maximum vorliegt; von dort aus fällt also $F_{\lambda,0} : F_{\lambda,\lambda}$ nach beiden Seiten ab. Der Wert dieses Verhältnisses bei $\lambda = \frac{3}{2}$ beträgt 1,03, und sowohl für $\lambda = 1$ als auch für $\lambda = 2$ ergibt Formel (16) den Wert des Verhältnisses $= 1$.

Somit ist bei keinem ganzzahligem λ das Verhältnis der Formel (16) > 1, und bei λ zwischen 1 und 2 nur um sehr wenig. Man kann deshalb davon ausgehen, dass bei jedem λ die Differenz $F_{\lambda,m}$ ihren höchsten Wert für $m = \lambda$ annimmt und sich zur Bestimmung des Höchstbetrages won $F_{\lambda,m}$ auf diesen Fall $m = \lambda$ beschränken.

Dann ist nach Formel (13), in der $m = \lambda$ gesetzt wird:

$$F_{\lambda,\lambda} = e^{-\lambda}\frac{\lambda^\lambda}{\lambda!}\,\frac{\lambda}{2s}, \text{ und da}$$

$$e^{-\lambda}\frac{\lambda^\lambda}{\lambda!} < \sqrt{\frac{1}{2\pi\lambda}}, \text{ so wird}$$

$$F_{\lambda,\lambda} < \frac{1}{2s}\sqrt{\frac{\lambda}{2\pi}}. \qquad (17)$$

Nun sei die Differenz zwischen dem exakten Werte $C_{\lambda,m}$ nnd der zweiten Näherung $B_{\lambda,m}$ betrachtet, um den Höchst-

betrag ihres absoluten Wertes zu ermitteln. Diese Differenz, die $H_{\lambda,m}$ genannt sei, ist gemäss den Formeln (1) und (10):

$$H_{\lambda,m} = \binom{s}{m}\left(\frac{\lambda}{s}\right)^m\left(1-\frac{\lambda}{s}\right)^{s-m} - \tag{18}$$

$$e^{-\lambda}\frac{\lambda^m}{m!}\left[1-\frac{(m-\lambda)^2-m}{2s}\right].$$

Wenn man hierin $\left(1-\frac{\lambda}{s}\right)^s$ durch den Näherungswert $e^{-\lambda}\left(1-\frac{\lambda^2}{2s}+\frac{3\lambda^4-8\lambda^3}{24s^2}\right)$ und den Zähler des Binomialkoeffizienten $\binom{s}{m}$ durch den Näherungswert $\left(s-\frac{m-1}{2}\right)^m$ ersetzt und weiterhin konsequent alle Glieder vernachlässigt, die dritte oder höhere Potenzen von s im Nenner haben, so findet man schliesslich näherungsweise:

$$H_{\lambda,m} = e^{-\lambda}\frac{\lambda^m}{m!} \tag{19}$$

$$\frac{3m^4-(12\lambda+9)m^3+(18\lambda^2+12\lambda+9)m^2-(12\lambda^3-6\lambda^2+3)m+3\lambda^4-8\lambda^3}{24s^2}$$

Ein Versuch, diesen ganzen Ausdruck nach m zu differenzieren und das Ergebnis $=0$ zu setzen, um so die Extremwerte von $H_{\lambda,m}$ unmittelbar zu bestimmen, führt zu einer transzendenten Gleichung von so unübersichtlicher Art, dass hieraus kein allgemeiner Einblick in das Wesen von $H_{\lambda,m}$ gewonnen werden kann. Es dürfte daher der beste Ausweg sein, den Ausdruck in (19) in zwei Faktoren zu zerlegen, um diese getrennt zu betrachten. Da der erste Faktor $= A_{\lambda,m}$, bleibt nur noch der Zähler des folgenden Bruches übrig, der $K_{\lambda,m}$ genannt sei, so dass also

$$H_{\lambda,m} = \frac{A_{\lambda,m}\,K_{\lambda,m}}{24s^2}.$$

Was zunächst $A_{\lambda,m}$ betrifft, so kann man dieses durch die Formel von Stirling ausdrücken, wobei die Korrekturglieder vernachlässigt werden. Wenn man diesen Ausdruck dann nach m differenziert, ergibt sich :

$$\frac{d\,A_{\lambda,m}}{d\,m} = \frac{e^{-\lambda}\,(e\,\lambda)^m}{m^m\,\sqrt{2\,\pi\,m}}\left(ln\,\lambda - ln\,m - \frac{1}{2\,m}\right). \qquad (20)$$

Der Bruch vor der Klammer verschwindet nur für $m = \infty$, in welchem Falle auch $A_{\lambda,m} = 0$ wird. Innerhalb der Klammer kann man $ln\,\lambda - ln\,m$ als Reihe entwickeln, und wenn man sich hierbei auf ein Glied beschränkt, so ergibt die Gleichung $ln\,\lambda - - ln\,m - \frac{1}{2\,m} = 0$ die einige Auflösung

$$m = \lambda - \frac{1}{2}\,.$$

Dies entspricht genau dem bekannten Verlaufe von $A_{\lambda,m}$, das bei ganzzahligem λ zwei gleiche Höchstwerte bei $m = \lambda - 1$ und $m = \lambda$ aufweist, während bei nicht ganzzahligem λ das Maximum bei demjenigen m erreicht wird, das zwischen $\lambda - 1$ und λ liegt.

Dann bleibt das in (19) auftretende Polynom zu betrachten, das als $K_{\lambda,m}$ bezeichnet wurde. Wenn man $\frac{d\,K_{\lambda,m}}{d\,m} = 0$ setzt, erhält man zur Bestimmung der Extremwerte von $K_{\lambda,m}$ die Gleichung :

$$m^3 - \frac{12\,\lambda + 9}{4}\,m^2 + \frac{6\,\lambda^2 + 4\,\lambda + 3}{2}\,m - \frac{4\lambda^3 - 2\lambda^2 + 1}{4} = 0. \qquad (21)$$

Die Diskriminante dieser Gleichung ist immer negativ, ausgenommen nur den in diesem Zusammenhange völlig bedeutungslosen Fall, dass $\lambda > 33{,}9$ wäre. Die Gleichung (21) hat also drei reelle Lösungen, die nur auf trigonometrischem Wege ermittelt werden können. Für einige Werte von λ sind diese Wurzeln der Gleichung (21) nachstehend zusammengestellt :

λ	m_1	m_2	m_3
1	0.128	1.715	3.407
2	0.446	2.804	5.000
3	0.905	3.901	6.444
4	1.441	5.000	7.809
5	2.026	6.101	9.124
6	2.646	7.203	10.401

Welchem von diesen je drei Werten das absolute Maximum von $K_{\lambda,m}$ entspricht, lässt sich nicht a priori entscheiden. Indessen kommt es auf dieses Maximum auch nicht an, sondern auf dasjenige des Produktes $A_{\lambda,m}\ K_{\lambda,m}$. Für dieses liegt der Gedanke nahe, dass es sich dort befindet, wo ein Extremwert von $K_{\lambda,m}$ möglichst nahe an dem Maximum von $A_{\lambda,m}$ liegt. Dies gilt für m_2, ausgenommen die kleinsten Werte von λ, wo m_1 in kleinerem Abstande von $\lambda - \frac{1}{2}$ liegt; hierbei ist jedoch der asymmetrische Verlauf von $A_{\lambda,m}$ zu berücksichtigen, durch den ein Wert von $A_{\lambda,m}$ mit $m > \lambda - \frac{1}{2}$ grösser sein kann als einer mit $m < \lambda - \frac{1}{2}$, wenn auch letztgenannter ein m hat, das näher an $\lambda - \frac{1}{2}$ liegt als das des erstgenannten.

Diese Ueberlegung führt, wenn sie auch nicht zwingend ist, dazu, das absolute Maximum von $H_{\lambda,m}$ bei einem m zu suchen, das zwischen $\lambda - \frac{1}{2}$ und obigem m_2 liegt, d.h. in der Nähe von λ. Die Tabelle bestätigt empirisch, dass bei allen ganzzahligen λ das Maximum des absoluten Wertes von $H_{\lambda,m}$ bei $m = \lambda$ liegt.

Für diesen Fall $m = \lambda$ lässt sich der tatsächliche Wert von $H_{\lambda,m}$ genauer berechnen. Zu diesem Zwecke setzt man in Formel (18) $m = \lambda$, erhält also:

$$H_{\lambda,\lambda} \binom{s}{\lambda} \left(\frac{\lambda}{s}\right)^{\lambda} \left(1 - \frac{\lambda}{s}\right)^{s-\lambda} - e^{-\lambda} \frac{\lambda^{\lambda}}{\lambda!} \left(1 + \frac{\lambda}{2s}\right). \quad (22)$$

Dies ergibt

$$H_{\lambda,\lambda} = \binom{s}{\lambda}\left(\frac{\lambda}{s}\right)^{\lambda}\left(1-\frac{\lambda}{s}\right)^{s}\left(\frac{s}{s-\lambda}\right)^{\lambda} - e^{-\lambda}\frac{\lambda^{\lambda}}{\lambda!}\left(1+\frac{\lambda}{2s}\right);$$

ersetzt man nun wie früher $\left(1-\frac{\lambda}{s}\right)^{s}$ durch den Näherungswert $e^{-\lambda}\left(1-\frac{\lambda^2}{2s}+\frac{3\lambda^4-8\lambda^3}{24s^2}\right)$, so wird aus der letzten Form von $H_{\lambda,\lambda}$:

$$H_{\lambda,\lambda} = e^{-\lambda}\frac{\lambda^{\lambda}}{\lambda!}\left[\frac{s(s-1)(s-2)\ldots\ldots(s-\lambda+1)}{(s-\lambda)^{\lambda}}\left(1-\right.\right.$$

$$\left.\left.-\frac{\lambda^2}{2s}+\frac{3\lambda^4-8\lambda^3}{24s^2}\right)-1-\frac{\lambda}{2s}\right].$$

Nun löst man den ersten Bruch innerhalb der eckigen Klammern in λ einzelne Brüche auf, d.h. man gibt ihm die Form

$$\left(1+\frac{\lambda}{s-\lambda}\right)\left(1+\frac{\lambda-1}{s-\lambda}\right)\left(1+\frac{\lambda-2}{s-\lambda}\right)\ldots\ldots\left(1+\frac{1}{s-\lambda}\right).$$

Wenn man jetzt das Produkt dieser λ Faktoren bildet und dabei genau wie bei der näherungsweisen Darstellung von $\left(1-\frac{\lambda}{s}\right)^{s}$ alle Glieder vernachlässigt, die im Nenner dritte oder höhere Potenzen von s bezw. von $s-\lambda$ enthalten, dann ergibt sich:

$$H_{\lambda,\lambda} = e^{-\lambda}\frac{\lambda^{\lambda}}{\lambda!}\left\{\left[1+\frac{\lambda(\lambda+1)}{2(s-\lambda)}+\frac{(\lambda-1)\lambda(\lambda+1)(3\lambda+2)}{24(s-\lambda)^2}\right]\right.$$

$$\left.\left(1-\frac{\lambda^2}{2s}+\frac{3\lambda^4-8\lambda^3}{24s^2}\right)-1-\frac{\lambda}{2s}\right\}.$$

Beim Multiplizieren der beiden Ausdrücke in Klammern werden wieder die Glieder mit höheren als zweiten Potenzen won s oder $s-\lambda$ im Nenner vernachlässigt. Wird das Ergebnis dann auf gemeinsamen Nenner gebracht und soweit möglich vereinfacht, dann wird es zu:

$$H_{\lambda,\lambda_1} = e^{-\lambda}\frac{\lambda^\lambda}{\lambda!}\;\frac{9\lambda^2 s^2 - 2\lambda s^2 + 10\lambda^4 s - 12\lambda^3 s + 3\lambda^6 - 8\lambda^5}{24 s^2 (s-\lambda)^2}.$$

Da im Zähler gegenüber den Gliedern mit s^2 diejenigen vernachlässigt werden dürfen, die s nur in der ersten Potenz oder garnicht enthalten, ist das endgültige Näherungsergebnis:

$$H_{\lambda,\lambda} = e^{-\lambda}\frac{\lambda^\lambda}{\lambda!}\;\frac{9\lambda^2 - 2\lambda}{24 (s-\lambda)^2}. \tag{23}$$

Bei der Ableitung von (16) wurde schon darauf hingewiesen, dass $e^{-\lambda}\frac{\lambda^\lambda}{\lambda!} < \frac{1}{\sqrt{2\pi\lambda}}$, und somit kommt man von (23) aus zu:

$$H_{\lambda,\lambda} < \frac{3\lambda}{8 (s-\lambda)^2}\sqrt{\frac{\lambda}{2\pi}}. \tag{24}$$

Hiermit kann auch der Höchstbetrag bequem berechnet werden, den der absolute Wert von $C_{\lambda,m} - A_{\lambda,m}$ annimmt; er sei $G_{\lambda,m}$ genannt. Da $C_{\lambda,m} - B_{\lambda,m}$ und $B_{\lambda,m} - A_{\lambda,m}$ ihren höchsten Wert bei $m = \lambda$ erreichen und beide Differenzen an dieser Stelle gleiches Vorzeichen haben, ist $C_{\lambda,\lambda} - A_{\lambda,\lambda} = (C_{\lambda,\lambda} - B_{\lambda,\lambda}) + (B_{\lambda,\lambda} - A_{\lambda,\lambda})$, d.h. mit den eingeführten Bezeichnungen:

$$G_{\lambda,\lambda} = F_{\lambda,\lambda} + H_{\lambda,\lambda}. \tag{25}$$

Gemäss den Formeln (17) und (24) ist also

$$G_{\lambda,\lambda} < \frac{1}{2s}\sqrt{\frac{\lambda}{2\pi}} + \frac{3\lambda}{8 (s-\lambda)^2}\sqrt{\frac{\lambda}{2\pi}},$$

und da die rechte Seite dieser Ungleichung auch geschrieben werden kann

$$\sqrt{\frac{\lambda}{2\pi}}\;\frac{4s^2 - 4s\lambda - \lambda(s - 4\lambda)}{8s (s-\lambda)^2},$$

so genügt die als selbstverständlich anzusehende Bedingung $s > 4\lambda$, um zu dem Ergebnis zu kommen:

$$G_{\lambda,\lambda} < \frac{1}{2(s-\lambda)} \sqrt{\frac{\lambda}{2\pi}} . \qquad (26)$$

Die im vorhergehenden zunächst empirisch angenommene Tatsache, dass $H_{\lambda,m}$ gegenüber $G_{\lambda,m}$ unbedeutend ist, lässt sich jetzt aus (24) und (26) deutlich erkennen. Dass in beiden Formeln ein Ungleichheitszeichen steht, ist kein entscheidendes Hindernis, weil der Grad der Ungleichheit in beiden Fällen ähnlich ist, sodass wenigstens der absolute Höchstbetrag dieser Differenzen seiner Grössenordnung nach mittels der Formeln (24) und (26) verglichen werden kann; also für $m = \lambda$ sei:

$$H_{\lambda,\lambda} : G_{\lambda,\lambda} = 3\lambda : 4(s-\lambda) . \qquad (27)$$

Da von der Voraussetzung ausgegangen wird, dass s weit grösser als λ ist (wenn auch nicht « sehr gross » gegenüber λ wie nach der Ableitung der Formel von *Poisson*), ergibt sich aus (27), dass die grösste Abweichung $H_{\lambda,m}$ weit kleiner als die grösste Abweichung $G_{\lambda,m}$ ist.

Bisher war davon ausgegangen worden, dass man zur Vereinfachung der Berechnungen λ als eine ganze Zahl betrachte, sodass ein m vorkommen könne, für das $m = \lambda$ gelte. Nun ist noch zu versuchen, ob man die Beweisführung auch durch Verzicht auf diese Annahme allgemeiner gestalten kann. Dies soll wenigstens für die Funktion $F_{\lambda,m}$ durchgeführt werden.

Wegen der Stetigkeit der in Betracht kommenden Funktionen kann vorausgesetzt werden, $F_{\lambda,m}$ erreiche seinen absoluten Höchstbetrag, wenn λ keine ganze Zahl ist, bei einem der beiden m zwischen denen λ liegt. Es sei $\lambda = c + h$, wobei c die nächste unter λ liegende ganze Zahl bedeutet, sodass h also ein echter Bruch ist, dann sind $F_{\lambda,c}$ und $F_{\lambda,c+1}$ zu prüfen.

Bezeichnet man, da es nur auf den absoluten Wert ankommt, unter Umkehrung des Vorzeichens in (13) die Differenz $B_{\lambda,m} - A_{\lambda,m}$ mit $F_{\lambda,m}$, so ergibt sich aus (13);

$$F_{\lambda,c} = e^{-(c+h)} \frac{(c+h)^c}{c!} \frac{c-h^2}{2s} . \qquad (28)$$

Wenn man dann wieder $c!$ nach Stirling unter Vernachlässigung der Korrekturglieder ausdrückt und beachtet, dass

$(c+h)^c < c^c e^h$, so findet man $F_{\lambda,c} < \frac{c - h^2}{2s\sqrt{2\pi c}}$ und da $\frac{c-h^2}{2s\sqrt{2\pi c}} <$

$< \frac{c+h}{2s\sqrt{2\pi(c+h)}}$, also

$$F_{\lambda,c} < \frac{1}{2s}\sqrt{\frac{\lambda}{2\pi}}. \tag{29}$$

Analog ist

$$F_{\lambda,c+1} = e^{-(c+h)}\frac{(c+h)^{c+1}}{(c+1)!}\cdot\frac{c^2+2h-h^2}{2s}. \tag{30}$$

Entwickelt man diesen Ausdruck auf die gleiche Weise und vergleicht man das Ergebnis mit $\frac{1}{2s}\sqrt{\frac{\lambda}{2\pi}}$, so ergibt sich:

$$F_{\lambda,c+1} : \frac{1}{2s}\sqrt{\frac{\lambda}{2\pi}} = \left(1+\frac{h-h^2}{c+h}\right)\left(1-\frac{1-h}{c+1}\right)^{c+1}\sqrt{\frac{c+h}{c+1}}. \tag{31}$$

Auf der rechten Seite von (31) kann man von der Wurzel absehen, da diese auf jeden Fall <1 ist. Das verbleibende Produkt der beiden Klammerausdrücke muss, wenn $c=0$ gesetzt wird, <1 sein. Da nun mit wachsendem c der erste Klammerausdruck stärker abnimmt, als der zweite zunimmt, ist also für jedes c die rechte Seite von (31) <1, und somit:

$$F_{\lambda,c+1} < \frac{1}{2s}\sqrt{\frac{\lambda}{2\pi}}. \tag{32}$$

Daher kann $F_{\lambda,m}$ für jedes λ und jedes m entsprechend (17) begrenzt werden mit:

$$F_{\lambda,m} < \frac{1}{2s}\sqrt{\frac{\lambda}{2\pi}}. \tag{33}$$

Da sich aus (27) ergab, dass in $G_{\lambda,\lambda}$ der überwiegende Teil durch $F_{\lambda,\lambda}$ gebildet wird, lässt sich — wenn auch ohne strengen Beweis — annehmen, dass Formel (26) für $G_{\lambda,m}$ auch gültig

bleibt, wenn λ nicht ganzzahlig ist, also der Höchstwert bei einem m erreicht wird, das nicht genau mit λ zusammenfällt.

Für $H_{\lambda,m}$ lässt sich bei nicht ganzzahligem λ kein allgemeiner Beweis dafür erbringen, dass die der Formel (24) entsprechende Ungleichung $H_{\lambda,m} < \frac{3\lambda}{8\sqrt{(s-\lambda)^2}}\sqrt{\frac{\lambda}{2\pi}}$ auch dann immer gültig bleibt; man muss sich hier mit dem empirisch gewonnenen Ergebnis begnügen.

Dieses ist in der nachfolgenden Tabelle mitenthalten, in der für einige λ, darunter auch nicht ganzzahlige, und für verschiedene s die tatsächlich errechneten absolut höchsten Werte von $G_{\lambda,m}$ und $H_{\lambda,m}$ angegeben und mit den theoretischen Höchstwerten gemäss (26) bzw. (24) verglichen sind.

Auf Grund der Formeln (26) bzw. (24) lässt sich schliesslich die Kernfrage dieser Untersuchung beantworten; sie lautet:

Wie gross muss s bei gegebenem λ oder q mindestens sein, damit in der Verteilung der $A_{\lambda,m}$ bzw. $B_{\lambda,m}$ der absolute Höchstbetrag der Abweichung vom exakten Werte $C_{\lambda,m}$ einen festgesetzten Wert ρ nicht überschreite?

Die gesuchte Mindestzahl für s als Funktion von ρ sei, je nachdem ob sie für $A_{\lambda,m}$ oder für $B_{\lambda,m}$ gelten soll, mit $s_{\varrho,A}$ bzw. $s_{\varrho,B}$ bezeichnet. Dann folgt aus (26) bzw. (24) unmittelbar:

$$\frac{1}{2(s_{\varrho,A}-\lambda)}\sqrt{\frac{\lambda}{2\pi}} < \rho \qquad \text{bzw.} \qquad \frac{3\lambda}{8(s_{\varrho,B}-\lambda)^2}\sqrt{\frac{\lambda}{2\pi}} < \rho\,.$$

Hieraus ergibt sich, wenn man $\sqrt{\frac{1}{2\pi}}$ zur Abkürzung β nennt:

$$s_{\varrho,A} > \frac{\beta}{2\rho}\sqrt{\lambda} + \lambda \tag{34}$$

bzw.

$$s_{\varrho,B} > \sqrt{\frac{3\beta}{8\rho}}\,\lambda^{\frac{3}{4}} + \lambda\,. \tag{35}$$

Als Beispiel sei im nachstehenden für ρ der Wert 0,001 gewählt. Mit diesem erhält man folgende Tabelle:

s	λ	m	$10^9 A_{\lambda,m}$	$10^9 B_{\lambda,m}$	$10^9 C_{\lambda,m}$	$10^9 G_{\lambda,m}$	$10^9 \frac{1}{2(s-\lambda)}\sqrt{\frac{\lambda}{2\pi}}$	$10^9 H_{\lambda,m}$	$10^9 \frac{3\lambda}{8(s-\lambda)^2}\sqrt{\frac{\lambda}{2\pi}}$
10	1	1	367 879 441	386 273 413	387 420 489	19 541 048	22 163 460	1 147 076	1 846 955
25	1	1	367 879 441	375 237 030	375 413 247	7 533 806	8 311 298	176 217	259 728
	2	2	270 670 566	281 497 389	282 111 804	11 441 238	12 264 991	614 415	799 891
50	0.5	1	303 265 330	305 539 820	305 558 620	2 293 290	2 849 442	18 800	21 587
	1	1	367 879 441	371 558 235	371 601 714	3 722 273	4 070 840	43 479	62 309
	1.5	2	251 021 430	255 414 305	255 518 202	4 496 772	5 037 139	103 897	116 842
	2	2	270 670 566	276 083 977	276 232 807	5 562 241	5 876 975	148 830	183 655
	3	3	224 041 808	230 763 062	231 056 900	7 015 092	7 350 939	293 838	351 907
	4	4	195 366 815	203 181 488	203 654 469	8 287 654	8 672 658	472 981	565 608
	5	5	175 467 370	184 240 738	184 924 601	9 457 231	9 911 801	683 863	825 983
100	0.5	1	303 265 330	304 402 575	304 407 254	1 141 924	1 417 562	4 679	5 343
	1	1	367 879 441	369 718 838	369 729 638	1 850 197	2 014 860	10 800	15 264
	1.5	2	251 021 430	253 217 868	253 243 499	2 222 069	2 480 216	25 631	28 327
	2	2	270 670 566	273 377 272	273 413 912	2 743 346	2 878 518	36 640	44 059
	3	3	224 041 808	227 402 435	227 474 127	3 432 319	3 561 795	71 692	82 619
	4	4	195 366 815	199 274 151	195 388 497	4 021 682	4 155 649	114 346	129 864
	5	5	175 467 370	179 854 054	180 017 827	4 550 457	4 695 063	163 773	185 331
	6	6	160 623 141	165 441 835	165 661 308	5 038 167	5 197 899	219 473	248 836
	7	7	149 002 780	154 217 877	154 498 993	5 496 213	5 674 742	281 116	320 348
	8	8	139 586 532	145 169 993	145 518 475	5 931 943	6 132 495	348 482	399 945
	9	9	131 755 640	137 684 644	138 106 057	6 350 417	6 575 972	421 413	487 778
	10	10	125 110 036	131 365 538	131 865 154	6 755 118	7 008 701	499 616	584 058
250	1	1	367 879 441	368 615 200	368 616 924	737 483	801 089	1 724	2 413
	2	2	270 670 566	271 753 248	271 759 058	1 088 492	1 137 479	5 810	6 880
	3	3	224 041 808	225 386 059	225 397 291	1 355 483	1 398 762	11 232	12 742
	5	5	175 467 370	177 222 044	177 247 604	1 780 234	1 820 535	25 560	27 865
	10	10	125 110 036	127 612 237	127 688 122	2 578 086	2 628 263	75 885	82 133
	15	15	102 435 862	105 508 938	105 652 276	3 216 414	3 287 440	143 338	157 377
	20	20	88 835 334	92 388 747	92 614 509	3 779 175	3 878 531	225 762	252 948

λ	$s_{\varrho,A}$	$s_{\varrho,B}$
0.5	141.5	7.8
1.0	200.5	13.2
1.5	245.8	18.1
2.0	284.1	22.6
2.5	317.9	26.8
3.0	348.5	30.9
3.5	376.7	34.8
4.0	402.9	38.6
4.5	427.6	42.3
5.0	451.0	45.9
5.5	473.3	49.4
6.0	494.6	52.9
6.5	515.1	56.3
7.0	534.8	59.6
7.5	553.8	62.9
8.0	572.2	66.2
8.5	590.1	69.4
9.0	607.4	72.6
9.5	624.3	75.7
10.0	640.8	78.8

Will man diese Tabelle für ein anderes ρ anwenden, so kann man berücksichtigen, dass in (34) allgemein und in (35) nur mit Ausnahme ganz kleiner Werte von λ das Zusatzglied $+\lambda$ gegenüber dem ersten Gliede bedeutungslos ist, so dass man einfach auf ein anderes ρ umrechnen kann, indem man die Tabellenwerte von $s_{\varrho,A}$ im umgekehrten Verhältnis von ρ und diejenigen von $s_{\varrho,B}$ im ungekehrten Verhältnis von $\sqrt{\rho}$ abändert.

Da jedoch die ursprünglich gegebene Grösse nicht λ, sondern q ist, empfiehlt es sich, die Mindestwerte $s_{\varrho,A}$ und $s_{\varrho,B}$ auch als Funktionen von q zu betrachten. Auf Grund der eingangs gegebenen Definition $s\,q = \lambda$ folgt aus (34) bzw. (35):

$$s_{\varrho,A} > \frac{q}{8\,p^2\,\pi\,\rho^2} \tag{36}$$

bzw.

$$s_{\varrho,B} > \frac{9\,q^2}{128\,p^4\,\pi\,\rho^2}\,. \tag{37}$$

Als Beispiel sei wieder $\rho = 0{,}001$ angenommen, und für diesen Fall ergibt sich folgende Tabelle:

q	$s_{\varrho,A}$	$s_{\varrho,B}$
0.01	406.0	0.0
0.02	828.6	0.2
0.03	1268.6	0.7
0.04	1726.9	1.7
0.05	2204.4	3.4
0.06	2701.8	6.2
0.07	3220.3	10.3
0.08	3760.8	16.0
0.09	4324.3	23.8
0.10	4912.2	34.1

Bei dieser Tabelle ist die Umrechnung auf irgendein anderes ϱ besonders einfach, weil man die Zahlen in beiden Spalten nur im umgekehrten Verhältnis von ϱ^2 abzuändern hat. Das Verhältnis von $s_{\varrho,A} : s_{\varrho,B}$ für ein gegebenes q ist aber von der Wahl der Genauigkeitsgrenze ϱ unabhängig. Daher zeigt diese letzte Tabelle besonders deutlich, wie sehr der Anwendbarkeitsbereich von $A_{\lambda,m}$ gemäss (1) erweitert wird, wenn man dieses zu $B_{\lambda,m}$ gemäss (10) ergänzt.

Für die Fälle $\lambda = 0{,}5$, $\lambda = 2$, $\lambda = 5$ und $\lambda = 10$ führt v. *Bortkiewicz* Beispiele vor, [4] wie sich die Abweichungen der im vorstehenden als $A_{\lambda,m}$ bezeichneten Näherungswerte von den $C_{\lambda,m}$ bei einigen Werten von s verhalten; als kleinstes s wählt er hierbei 1000. Da andererseits in seinen praktischen Beispielen [5] λ nie grösser als 5 ausfällt, genügt es, die Ergebnisse für diesen Fall zu betrachten. Dann ergibt sich aus (26), dass die absolut grösste Abweichung $< 0{,}00045$ sein muss; nach der exakten Berechnung durch v. *Bortkiewicz* beträgt sie 0,00044. Geht man im gleichen Falle aber von der Formel $A_{\lambda,m}$ zu der Formel $B_{\lambda,m}$ über, dann wird nach (24) die absolut grösste Abweichung $< 0{,}0000017$.

Geht man nicht von festen λ, sondern von festen q aus, so ergibt sich aus der letzten Tabelle, dass für ein q von 0,07, also im vorliegenden Zusammenhange einen verhältnismässig hohen Wert, zur Erzielung einer Genauigkeit von 0,001 nur ungefähr 10 Beobachtungen nötig sind, wenn man mit $B_{\lambda,m}$ rechnet, während bei $A_{\lambda,m}$ die Mindestzahl der Beobachtungen zur Gewin-

---

[4] l. c., p. 11, 12.
[5] l. c., p. 17 - 25.

nung einer gleichen Genauigkeit mehr als 3200 beträgt. Bei kleineren q, wie sie nach der Formulierung des Gesetzes der kleinen Zahlen die Regel sein müssen, verändert sich gemäss (36) und (37) das Verhältnis noch mehr zu Gunsten von $B_{\lambda,m}$ Dies wird schematisch deutlich, wenn man in beiden Formeln den Faktor p^2 bzw. p^4 vernachlässigt ; dann wird $s_{Q,B} : s_{Q,A} = \frac{9}{16} q^2 : 1$, und dieses Verhältnis nimmt bei sinkendem q schnell ab.

Zusammenfassung.

Die Wahrscheinlichkeit q für das Eintreten eines Ereignisses in s Beobachtungen sei klein, etwa $q < 0{,}1$; der Erwartungswert $s\,q$ werde mit λ bezeichnet. Die Wahrscheinlichkeit dafür, dass das Ereignis in den s Beobachtungen gerade m mal eintritt, sei auf drei Arten berechnet, und zwar sei

$$A_{\lambda,m} = e^{-\lambda} \frac{\lambda^m}{m!}$$ (die von v. *Bortkiewicz* eingeführte Formel),

$$B_{\lambda,m} = A_{\lambda,m} \left[1 - \frac{(m-\lambda)^2 - m}{2\,s} \right]$$ (die hier vorgeschlagene Korrektur),

$$C_{\lambda,m} = \binom{s}{m} \left(\frac{\lambda}{s}\right)^m \left(1 - \frac{\lambda}{s}\right)^{s-m}$$ (die exakte Formel).

Dann sind die absoluten Höchstabweichungen der beiden Näherungsformeln von der exakten Formel gegeben durch :

$$| A_{\lambda,m} - C_{\lambda,m} | = G_{\lambda,m} < \frac{1}{2\,(s-\lambda)} \sqrt{\frac{\lambda}{2\pi}}$$

und

$$| B_{\lambda,m} - C_{\lambda,m} | = H_{\lambda,m} < \frac{3\,\lambda}{8\,(s-\lambda)^2} \sqrt{\frac{\lambda}{2\pi}} ,$$

so dass näherungsweise $H_{\lambda,m} : G_{\lambda,m} = 3\,\lambda : 4\,(s-\lambda)$.

Soll bei gegebenem q der grösste Fehler durch Gebrauch der Näherungsformeln einen Betrag ρ nicht überschreiten, so lautet die Bedingung für $A_{\lambda,m}$:

$$s > \frac{q}{8\,p^2\,\pi\,\rho^2}$$

und für $B_{\lambda,m}$

$$s > \frac{9\,q^3}{128\,p^4\,\pi\,\rho^2}\,.$$

Wegen der vorausgesetzten Kleinheit von q erlaubt also die Formel $B_{\lambda,m}$ bis auf weit kleinere Beobachtungszahlen herabzugehen als $A_{\lambda,m}$. Ist z. B. $q = 0{,}07$, so genügen schon ungefähr 10 Beobachtungen, um den Fehler durchweg kleiner als 0,001 und ungefähr 1000 Beobachtungen, um ihn durchweg kleiner als 0,0001 zu halten.

Da die Berechnung von $B_{\lambda,m}$ sehr einfach ist, ist die Anwendung dieser Nährungsformel bei mässig grossem s (etwa unter 1000) zu empfehlen.